Les modèles Bloom pour le traitement automatique de la langue française

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

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

Rachel Bawden [2] **Hatim Bourfoune** [1] **Bertand Cabot** [1] **Nathan Cassereau** [1]
Pierre Cornette [1] **Marco Naguib** [3] **Aurélie Névéol** [3] **François Yvon** [4]

[1] CNRS, IDRIS, 91 403, Orsay, France
[2] Inria Paris, 75 012, 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 prendre en charge de multiples analyses automatiques de textes, implique de développer en parallèle l'infrastructure requise pour évaluer ces modèles, en couvrant idéalement le plus de tâches possible. De nombreux ensembles de données de références ont ainsi été rassemblés pour la langue anglaise, permettant d'évaluer ces « gigas modèles » sous de multiples facettes. Il existe également des jeux de test multilingues, avec une couverture bien moindre, qui sont utilisés pour mesurer la capacité de ces modèles à traiter plusieurs langues. Dans cet article, nous présentons nos propres efforts pour assembler un ensemble d'évaluation multi-tâche pour le français, qui est ensuite utilisé pour évaluer les modèles de la famille `Bloom`. Les résultats présentés ici confirment et complémentent les principaux résultats d'évaluation de `Bloom` en anglais ; ils permettent de conclure que les performances obtenues en français et en anglais sont très voisines, et encore meilleures lorsque les amorces utilisés pour l'inférence en contexte sont rédigées dans la même langue que les textes soumis à l'analyse.

English: The development of very large language models, capable of performing a large range of automatic language processing tasks, simultaneously requires to develop the infrastructure needed to evaluate these models, ideally covering as many tasks as possible. Numerous benchmarks have already been compiled for the English language, making it possible to evaluate these large models from multiple angles. Several multilingual test sets are also available, with a much lesser coverage, which are used to measure the ability of these models to handle multiple languages. In this paper, we present our efforts to assemble a multi-task evaluation set for French, which is then used to evaluate models from the `Bloom` family. Our results confirm and complement the main evaluation results for `Bloom` in English; they allow us to conclude that the performances obtained in French and English are very similar, and even better when the prompts used at inference are written in the same language as the texts to analyze.

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 (TAL). 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., 2019b; Brown et al., 2020), débruitage (Lewis et al., 2020)) pour lesquelles des annotations naturelles sont immédiatement 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 (quelques milliers) (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 *demonstrations*, 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., 2019b). 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 pouraient aider à 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 permis l'entraînement d'un giga modèle de langue *multilingue* selon les principes de la science ouverte. Ce modèle, `Bloom`, existe en plusieurs tailles, la plus petite version comprenant 560 millions de paramètres, la plus grande 176 milliards. Il est présenté en détail dans (BigScience et al., 2022), des publications ultérieures documentant des aspects particuliers du modèle. Ainsi le coût carbone de l'apprentissage est discuté dans (Luccioni et al., 2023) et une dérivation du modèle par affinage multitâche et multilingue donnant lieu aux familles `mT0` et `Bloomz` est décrite dans (Muennighoff et al., 2022). Ce modèle sera présenté dans la section 2.

Ce modèle est doublement intéressant : (a) il est totalement ouvert et facilement disponible ; de surcroît tous les détails concernant son entraînement (y compris les corpus) et son exploitation 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 un mélange de documents qui accorde une large part au français (environ 15% des données d'apprentissage) et aux langues romanes (35% des données), et dans un autre genre, aux langages de programmation.

Le projet Bloume Le projet Bloume, qui s'est déroulé durant le premier semestre 2023, avait pour objectif d'évaluer le modèles `Bloom` et `Bloomz` sur des jeux de tests en langue française afin (a) d'établir des performances de référence sur un large évantail 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.

Contributions Dans cet article, nous présentons le protocole et l'environnement de test, ainsi que les évolutions de cet environnement qui ont été réalisées dans le cadre du projet. Nous présentons également les principaux résultats obtenus sur les tâches considérées (modélisation de la langue, classification, implication textuelle, réponse à des questions, reconnaissance d'entités, traduction, résumé automatique, étude des biais), et en discutons la portée et les limites de ces résultats bruts. L'ensemble des codes et des résultats d'évaluation est accessible en ligne à l'adresse `https://github.com/fyvo/EvaluerBloom`.

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 *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).

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

2 Bloom, un giga modèle de langue

2.1 Généralités

Le projet BigScience [3] 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 1, 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 [4] 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.

	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 1 – Les modèles de la famille Bloom.

L'apprentissage de ces modèles exploite le corpus ROOTS [5] (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 [6]. 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 (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 émanents d'institutions internationnales, encyclopédie, matériel pédagogique, sous-titres, etc), soit à des corpus collectés sur le web et rassemblés dans le corpus OSCAR (Ortiz Suárez et al., 2019).

2.1.1 Bloomz

Bloomz [7] (Muennighoff et al., 2022) est un modèle dérivé de Bloom par affinage, en poursuivant le processus d'apprentissage sur des séquences textuelles reproduisants 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 [8]. 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

3. https://bigscience.huggingface.co/
4. https://huggingface.co/bigscience/bloom
5. Consultable à https://huggingface.co/spaces/bigscience-data/roots-search.
6. Les langues officiellement couvertes par Bloom sont documentées dans la carte d'identité du modèle : https://huggingface.co/bigscience/bloom.
7. https://huggingface.co/bigscience/bloomz
8. Les données utilisées pour l'affinage constituent le corpus xP3 : voir https://huggingface.co/datasets/bigscience/xP3.

sémantique, résumé, simplification et traduction automatique pour citer les principales. Il est important de noter que cet affinage utilise certains de nos propres jeux de tests : 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 (voir section 3.6 pour ces deux tâches de génération). Pour ces jeux de test, les résultats de `Bloomz` surestiment les performances réelles de ce modèle.

2.2 Évaluer `Bloom` avec des données françaises

Pour les évaluations, nous réexploitons le cadre méthodologique d'évaluation utilisé dans le cadre du projet BigScience. Ce cadre permet d'évaluer les performances du modèle lorsqu'on l'interroge directement à l'aide d'amorces pouvant inclure des exemples de la tâche à accomplir. Par comparaison avec une approche reposant sur l'affinage, cette méthode est plus légère car elle ne demande pas de réapprentissage ; pour de nombreuses tâches elle conduit probablement à des performances moins bonnes que l'affinage. Deux composants logiciels sont nécessaires pour mettre en œuvre cette démarche.

Le premier, `promptsource` [9] (Sanh et al., 2022), permet de constuire des amorces et des instructions dédiées au traitement d'un jeu de données et d'une tâche particulière. La version utilisée contient des amorces pour plusieurs centaines de jeux de données standard du TAL, principalement rédigées en anglais. Nous avons étendu ces jeux d'amorces par des équivalents en français pour les tâches considérées dans nos tests.

Le second, `lmharness` [10] (Gao et al., 2021), permet de 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 démsonstrations, les paramètres du processus de génération, etc.). Chaque appel à `lmharness` produit l'ensemble des textes générés automatiquement pour compléter l'amorce, ainsi que les résultats d'évaluations aggrégés pour la, ou les, métriques associées à la tâche. Lorsque l'amorce inclut une ou plusieurs démonstrations (dans nos expériences, au plus une), le choix de l'exemple résulte d'un tirage aléatoire parmi les données de développement associées à la tâche [11].

Dans le cadre de cette étude, diverses contributions techniques à `lmharness` ont été réalisées, en particulier pour intégrer des tâches de reconnaissance d'entités nommées et les métriques associées. Ces contributions ont fait l'objet de mises à jour de ces deux composants logiciels [12].

3 Protocoles, tâches, données et métriques

3.1 Modélisation de la langue

`Bloom` est un modèle de langue causal entraîné sur la tâche de prédiction du prochain mot. Une première évaluation porte donc sur sa capacité à réaliser correctement cette tâche. Les principales métriques utilisées dans la littérature sur les modèles de langue sont la *perplexité*, utilisée classiquement en théorie de l'information et le *nombre de bits par octet* (*bits per byte*), introduite plus récemment pour comparer des modèles en faisant abstraction de la segmentation en mots ou sous-mots qu'ils utilisent pour calculer la probabilité d'une séquence. Les formulations de ces métriques sont rappelées ci-dessous :

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

$$\text{bits-par-octet} = CE(|B|) \tag{2}$$

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

Pour le calcul de (1), nous réinitialisons le contexte gauche après chaque phrase ; $|B|$ désigne le nombre total d'octets du texte $w_1 \dots w_T$ [13].

Ces mesures impliquent de choisir un corpus de textes approprié. Simoulin and Crabbé (2021) utilisent un corpus dérivé de la Wikipédia française à partir d'articles mis en avant pour leur qualité et

9. `https://github.com/bigscience-workshop/promptsource`.
10. `https://github.com/bigscience-workshop/lm-evaluation-harness/`.
11. La graine aléatoire est la même pour toutes les expériences.
12. Voir, de nouveau, `https://github.com/fyvo/EvaluerBloom`
13. Ce nombre dépend de l'encodage du texte. Nous utilisons UTF-8.

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

TABLE 2 – Corpus pour l'évaluation de la génération.

comprenant un 60 articles de test. Dans la mesure où ces articles font partie du corpus d'apprentissage de Bloom, nous utilisons à la place deux corpus comparables 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 présente l'avantage d'être multiparallèle, donc permettant des comparaisons entre langues et n'a pas été utilisé pour entraîner Bloom. Il présente toutefois des traces de la langue source – il faut donc considérer ce corpus comme du traductionnais (*translationese*). Pour contrôler cet effet, nous reproduisons la méthodologie de (Simoulin and Crabbé, 2021) en collectant le corpus wikitext-fr-2022, constitué de 60 articles récents (postérieurs à 2022) sélectionnés pour leur qualité dans la Wikipédia française [14]. Les statistiques de base de ces deux corpus sont dans le tableau 2.

Pour mémoire, les perplexités obtenues sont respectivement de 109,2 et 12,9 pour le petit (resp. grand) modèle GPT-FR (voir (Simoulin and Crabbé, 2021), tableau 4) [15].

3.2 Classification de textes

Pour la classification de textes, nous considérons la tâche Multilingual Amazon Reviews [16], qui propose plusieurs problèmes de prédiction du niveau d'appréciation d'un produit à partir d'un commentaire textuel. Dans sa version de base, elle consiste à assigner un score entre 1 et 5 à partir du commentaire ; des variantes de la tâche n'utilisent que le titre du commentaire, ou bien encore à la fois le titre et le corps du texte. Cette tâche fait partie de celles qui sont intégrées dans divers benchmarks comme FLUE (Le et al., 2020a) et BUFFET (Asai et al., 2023). La version de FLUE est toutefois plus simple, et ne distingue que les commentaires positifs (score au moins 4) et négatifs (scores au plus 2), les commentaires ayant reçu la note 3 étant supprimés. Avec cette restriction, pour cette tâche, les résultats de FlauBERT (Large) après affinage sont autour de 95% d'étiquettes correctement prédits (voir (Le et al., 2020a), tableau 3).

La version que nous considérons correspond au cadre standard d'évaluation pour les tâches de classification : le système est supposé avoir répondu correctement à l'amorce lorsque la réponse correcte est la plus probable parmi les alternatives possibles *listées dans l'amorce* (ici les chiffres de 1 à 5), et ceci même lorsque d'autres sorties seraient encore plus probables [17]. Les amorces utilisées s'apparentent toutes au modèle suivant :

> 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 ainsi construits, qui ne se distinguent que par l'information utilisée : soit le titre seul, comme dans l'exemple ci-dessus, soit le corps seul du commentaire, soit la conjonction du titre et du corps du commentaire [18]. Ces formulations reprennent les choix qui ont été faits pour exprimer les amorces en anglais par (Sanh et al., 2022) et permettent donc des comparaison entre langues. Nous n'utilisons que les données de test associées à cette tâche, soit 5 000 commentaires pour le français et autant pour l'anglais.

14. Le code utilisé est celui des auteurs https://github.com/AntoineSimoulin/gpt-fr. Chaque ligne de ce corpus correspond à un court paragraphe.
15. Le facteur de normalisation pour le calcul de la perplexité est le nombre de tokens identifiés lors de la segmentation du texte.
16. https://huggingface.co/datasets/amazon_reviews_multi/viewer/all_languages/test
17. Une excellente discussion des subtilités de l'évaluation des LLM et de la difficulté de comparer les scores est menée dans (Fournier et al., 2023).
18. Les formulations sont alors respectivement « Sur la base du commentaire... », « Sur la base du commentaire et de son titre ».

3.3 Équivalences sémantiques

La tâche de calcul d'équivalences sémantiques est une tâche ancienne, introduite dans (Monz and de Rijke, 2001; Dagan and Glickman, 2004) qui prend des formes [19] 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 : l'énoncé 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 ? Selon les formulations, la relation à identifier peut être posée comme étant symétrique, non symétrique, voire anti-symétrique. Grau and Gleize (2018) discute des différentes tâches d'implication qui sont usuellement considérées, ainsi que des difficultés qu'elles posent aux machines. Sous une forme ou une autre, cette tâche est intégrée dans les benchmarks standard en TAL tels que `GLUE` (Wang et al., 2019b) ou `SuperGLUE` (Wang et al., 2019a).

Sous l'impulsion initialement des évaluations proposées dans le « réseau d'excellence » Pascal, puis sous l'égide de la série d'ateliers SemEval, plusieurs jeux de données ont été développés, d'abord en anglais, puis dans d'autres langues. La tâche a également été étendue à un cadre complètement multilingue : A et B sont alors exprimées dans des langues différentes (Negri et al., 2012). Ces données multilingues sont souvent obtenues par traduction (automatique ou humaine) à partir de données monolingues.

Pour ce travail, nous avons choisi d'utiliser le jeu de test `XNLI` (Conneau et al., 2018), construit par traduction depuis l'anglais en 14 langues d'un petit ensemble de phrases extraites du corpus `MNLI` (Bowman et al., 2015; Williams et al., 2018). 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` [20] (Le et al., 2020a), et des résultats pour ces données sont disponibles dans plusieurs travaux en français. (Le et al., 2020a) rapporte ([tableau 5) des scores de correction [21] compris entre 76,9 et 85,2, les modèles `FlauBERT` atteignant respectivement 80,6 et 83,4 et `CamemBERT` 81,2.

`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 (voir références supra).

Pour réaliser les évaluations de la section 4.3, nous avons traduit en français les amorces de (Bach et al., 2022) (voir le tableau 3). Comme il est usuel, la métrique utilisée est la correction. Pour cette tâche, comme pour les autres tâches de classification, la réponse du système est jugée correcte lorsqu'elle est la plus probable *des réponses associées à l'amorce*. Chaque amorce du tableau 3 correspond à un ensemble de trois réponses possibles, formulées à chaque fois dans la même langue que l'amorce.

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

La reconnaissance des entités nommées (REN) consiste en l'identification et la classification des *entités nommées*, qui sont des mentions qui font référence à des entités du monde réel. On peut alors s'intéresser à des entités du domaine général (personnes, lieux, organisations), ou à des entités plus spécifiques à un domaine particulier comme le domaine clinique (maladie, symptôme, partie du corps) ou le domaine juridique (court, loi, délit). Pour le domaine général, nous utilisons le jeu de test de `WikiNER_fr` (Nothman et al., 2013) [22] qui comporte 13 410 exemples. Pour le domaine clinique, nous utilisons le jeu de test de `QuaeroFrenchMed` (Névéol et al., 2014) [23]. La figure 1 présente un extrait annoté d'un document EMEA (les entités sont annotées au niveau de chaque token, avec 2=CHEMICAL AND DRUGS, 4=DISORDER, 6=LIVING BEINGS, 10=PROCEDURE). Nous fixons la longueur maximum des prédictions à 64 tokens pour `WikiNER_fr` et 32 tokens pour `QuaeroFrenchMed`.

19. Et des dénominations multiples : « reconnaissance d'implications textuelles », « inférence en langue naturelle », et « détection de paraphrases », chacune recouvrant des tâches légèrement différentes.
20. `https://github.com/getalp/Flaubert/blob/master/flue/`
21. Définie comme le pourcentage de bonnes réponses parmi les sorties du modèle.
22. `https://huggingface.co/datasets/Jean-Baptiste/wikiner_fr`
23. `https://huggingface.co/datasets/meczifho/QuaeroFrenchMed`

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

(1) 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

FIGURE 1 – Exemple de phrase annotée tiré du corpus EMEA

Reconnaissance d'entités nommées générales Les annotations non anglaises de WikiNER, composé d'articles de Wikipédia, ont été 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 (en anglais *silver annotations*). 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).

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 4. Un exemple du jeu de données initiales (annotations au niveau de chaque token ; 1=lieu, 2=personne, 4=organisation) est donné dans l'exemple 2 et sa transformation en amorce de type LIST_PER est donnée dans l'exemple 3.

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

(2) Les poètes Joachim du Bellay et Pierre Ronsard sympathisent à
0 0 2 2 2 0 2 2 0 0
l' Université de Poitiers , avant de monter à Paris .
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 stratégie, 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 sur la base d'un petit ensemble de délimiteurs (retour chariot, point-virgule, et virgule [24]), (ii) la suppression d'apostrophes, guillemets et espaces dans chaque élément (pour éviter que les variations de ponctuation soient comptés comme des erreurs) ; (iii) le calcul du nombre de prédictions correctes sur la base des deux listes résultantes, sans tenir compte de l'ordre des éléments dans les deux listes, 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 `QuaeroFrenchMed` (Névéol et al., 2014) est un corpus composé de deux parties. La première partie, `EMEA` est une collection de 13 notices concernant des médicaments commercialisés en Europe, fournis par l'Agence Européenne des Médicaments. La seconde partie, `MEDLINE`, consiste en 2 500 titres d'articles scientifiques indexés dans la base de données bibliographique MEDLINE [25]. Ces deux parties sont annotées en 10 types d'entités nommées correspondant à des groupes sémantiques de l'UMLS (*Unified Medical Language System*) (Bodenreider and McCray, 2003) : symptômes et maladies, parties du corps, composants chimiques, êtres vivants, procédures médicales, physiologie humaine, phénomènes physiologiques, dispositifs médicaux, zones géographiques et objets. Ce corpus a été utilisé lors de la campagne d'évaluation CLEF eHealth et le meilleur système offrait une F-mesure de 0.749 sur la partie `EMEA` et de 0.698 sur partie `MEDLINE` (Névéol et al., 2016). Ces performances, obtenues avec un système symbolique, continuent de représenter l'état de l'art pour `QuaeroFrenchMed`.

Nous utilisons les deux stratégies d'évaluation décrites ci-dessus, en utilisant des amorces explicitant la nature du document (selon la partie du corpus considérées). Les amorces sont présentées dans le tableau 5. Nous utilisons les mêmes métriques que pour la reconnaissance d'entités nommées générales.

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

24. 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.

25. `http://pubmed.ncbi.nlm.nih.gov/`

3.5 Réponses aux questions

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 il assure que la réponse à la question posée figure toujours intégralement dans le passage présenté. PIAF ne représente que très imparfaitement la variété des tâches de réponses aux questions, dont on peut se faire une idée en consultant (Rogers et al., 2023). Un autre jeu de données pour la tâche de réponse aux questions en français est FQuAD (d'Hoffschmidt et al., 2020).

L'évaluation repose sur deux métriques classsiques : la première (EM, pour *Exact Match*) correspond au pourcentage de questions pour lesquelles la réponse fournie est exactement la réponse de référence ; la seconde (F1) combine précision et rappel au niveau des mots de la réponse et fournit donc une évaluation moins stricte des performances. L'évaluation porte sur l'intégralité des questions de PIAF 1.0, soit 3 835 couples de (question, réponse).

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 dans (Cattan et al., 2021) qui l'utilise pour comparer un ensemble de modèles de langue pour le français, ainsi que divers modèles multilingues. Dans cette étude, l'évaluation ne concerne qu'un ensemble restreint de questions ; les meilleurs résultats présentés sont obtenus en affinant le modèle XLM-R$_{large}$ (F1=73,2, EM=45.8).

Pour cette tâche nous considérons quatre amorces qui ont déjà été utilisées dans des études similaires, chaque amorce existant en français et en anglais. Un autre trait important de ces amorces est la position relative de l'instruction, du passage et de la question.

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 6 – 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.

3.6 Génération de textes

Comme dans l'article décrivant Bloom, nous considérons deux tâches de génération de textes : le résumé et la traduction automatique, en nous appuyant sur les mêmes jeux de données. Pour ces deux tâches, nous suivons (BigScience et al., 2022) et utilisons un modèle de génération glouton, en fixant à l'avance une longueur maximale de 512 *tokens* pour les textes générés.

3.6.1 Résumé automatique

Concernant le résumé, il s'agit de WIKILINGUA (Ladhak et al., 2020), un corpus de test multilingue construit par alignement semi-automatique de documents collectés sur le site WikiHow [26]. Nous utilisons la version réduite du jeu de test préparée par Gehrmann et al. (2022), qui contient 3 000 documents à résumer. Cette version est utilisée dans (BigScience et al., 2022), qui rapporte des résultats pour 9 langues, dont le français, avec des amorces en anglais. La principale nouveauté de ce travail est de reproduire ces résultats avec des amorces en français, qui s'inspirent au mieux de leurs équivalents anglais (tableau 7).

Nous calculons plusieurs métriques typiques du résumé automatique de la famille ROUGE (ROUGE-1, ROUGE-2, etc) (Lin, 2004). Nos discussions s'appuient sur la métrique ROUGE-2, qui est aussi

26. https://www.wikihow.com/

Langue	Nom	Amorce	Continuation
fr	summarize_above_fr	[Long texte] ¶¶Ecrivez un résumé du texte ci-dessus en français :	[résumé]
en	summarize_above_fr	[Long texte] ¶¶Write a summary of the text above in French :	[résumé]
fr	rephrase_fr	[Long texte] ¶¶Comment pourrait-on reformuler cela en français :	[résumé]
en	rephrase_fr	[Long texte] ¶¶How would you rephrase that in French ?	[résumé]
fr	tldr_fr	[Long texte] ¶¶Pour résumer, en français :	[résumé]
en	tldr_fr	[Long texte]¶¶TL ;DR in French :	[résumé]
fr	write_abstract_fr	D'abord, lisez le texte ci-dessous. ¶¶[Long texte] ¶¶Maintenant, s'il-vous-plait, écrivez un court résumé en français.r	[résumé]
en	write_abstract_fr	First, read the French article below. ¶¶[Long texte] ¶¶Now, please write a short abstract for it in French.	[résumé]
fr	article_summary_fr	Article en français :¶¶[Long texte] ¶¶Résumé en français.	[résumé]
en	article_summary_fr	Article in French : ¶¶[Long texte] ¶¶Summary in French.	[résumé]

TABLE 7 – Amorces utilisés pour la tâche de résumé automatique (corpus Wiki_lingua-fr). Les amorces existent systématiquement en deux langues (en, fr), avec des versions plus ou moins verbeuses.

utilisée par d'autres études sur le résumé automatique [27]. Cette métrique correspond simplement au recouvrement des bigrammes entre la référence et l'hypothèse : pour l'évaluer, nous donnerons les valeurs de précision, rappel et F-mesure.

3.6.2 Traduction automatique

La tâche de traduction automatique (TA) a été utilisée pour mettre en évidence les capacités multilingues de la plupart des GLM, depuis GPT-2 (Radford et al., 2019a) jusqu'à PALM (Chowdhery et al., 2022) et FALCON (Almazrouei et al., 2023). Dans la plupart des études, la traduction entre anglais et français fait partie des directions de traduction considérées, bien que cette paire de langues soit généralement considérée comme relativement simple. Le jeu de données provient le plus souvent de l'évaluation réalisée dans le cadre de l'atelier WMT 2014 (Bojar et al., 2014), en dépit des problèmes méthodologiques que pose l'utilisation d'un jeu de test déjà ancien, qui a déjà potentiellement été intégré dans les corpus d'apprentissage des grands modèles (Vilar et al., 2022) [28].

Langue	Nom	Amorce	Continuation
en	a_good_translation	Given the following source text (in L1) : [phrase source], a good L2 translation is :	[phrase cible]
en	translate_as	[phrase source] translates into L2 as :	[phrase cible]
en, fr	xglm	(L1 :) [phrase source] = L2 :	[phrase cible]
fr	fr_translate	Traduire (de L1) en L2 : [phrase source]	[phrase cible]
fr	fr_a_good_translation	Etant donné le texte source suivant (en L1) : [phrase source], une bonne traduction en L2 est :	[phrase cible]

TABLE 8 – Amorces utilisées pour la tâche de traduction automatique (corpus WMT'14). Toutes les amorces spécifient la langue cible (L2), mais le nom de la langue source n'est explicite que pour les amorces de la forme XXX-source+target ; pour les autres (XXX-target), elle n'est pas indiquée. Le amorce a_good_translation existe ainsi en deux versions par langue (avec ou sans langue source).

Afin de pouvoir réutiliser les résultats de Bawden and Yvon (2023) qui considèrent ce jeu de test, nous utilisons ces données, qui contiennent 3 003 phrases tirées de divers articles de presse [29]. Les amorces s'inspirent également des travaux antérieurs (voir le tableau 8 qui donne les 4 amorces pour chacune des langues).

27. L'implémentation de ROUGE est celle de http:github.com/google-research/google-research/rouge, commit f935042.

28. Nous avons vérifié que cela n'est pas le cas de ROOTS . Voir aussi sur les questions de méthode le billet très détaillé de Marie (2022).

29. Les phrases sont pour partie tirées d'articles en français (traduits en anglais), en anglais (traduits en français), ou bien des traductions depuis d'autres langues vers l'anglais et le français.

Nous évaluons les performances de traduction avec la métrique BLEU (Papineni et al., 2002) telle qu'implémentée par Post (2018) dans SacreBLEU [30].

3.7 Biais

Pour l'évaluation des biais, nous nous appuyons sur le corpus `CrowS-Pairs` proposé en anglais par (Nangia et al., 2020) puis corrigé et étendu au français par (Névéol et al., 2022). Ce corpus est composé de 1 610 paires de phrases qui explicitent des stéréotypes liés à neuf types de biais : race/couleur de peau, genre/identité de genre, statut socio-économique, profession, nationalité, âge, orientation sexuelle, apparence physique, handicap. Ce corpus repose sur le paradigme de *paire minimale* : chaque phrase est associée à une autre, qui ne diffère que d'un token représentant soit la catégorie sociale visée par le stéréotype ou une autre catégorie (la phrase est alors un anti-stéréotype). Nous présentons ci-dessous un exemple extrait du corpus, portant sur un stéréotype de genre (énoncé 1) et un anti-stéréotype associé (énoncé 2) :

1. ***Women*** *can't drive. / Les* ***femmes*** *ne savent pas conduire.*
2. ***Men*** *can't drive. / Les* ***hommes*** *ne savent pas conduire.*

L'évaluation, originellement prévue pour des modèles de langue masqués, cherche à déterminer si les modèles présentent une préférence systématique pour les énoncés stéréotypés. Le *metric score* de (Nangia et al., 2020) approxime la log-probabilité conditionnelle (*pseudo-log likelihood*) d'une sous-partie d'une phrase conditionnellement à l'autre partie. En comparant les scores de deux phrases d'une même paire, conditionnellement aux mots qui les distinguent, ces auteurs déterminent pour chaque paire si l'énoncé stéréotypique est plus ou moins probable que le contraste. Nous reprenons ici le protocole proposé par (BigScience et al., 2022) pour l'évaluation de modèles auto-regressif à l'aide d'une série d'amorces. Les amorces sont conçues pour éliciter du modèle évalué le choix d'une phrase de chaque paire afin de maximiser la probabilité et la vraisemblance des énoncés. La mesure d'exactitude obtenue pour un modèle non biaisé devrait ainsi être proche de 0,5, ce qui dénoterait une préférence aléatoire pour les stéréotypes ou les anti-stéréotypes.

Nous utilisons également le corpus `Shades` proposé par le groupe de travail sur l'évaluation des biais du collectif BigScience. Ce corpus s'appuie sur des stéréotypes de nationalité associés à des informations sur la perception des stéréotypes dans différentes cultures. Il s'agit d'une tentative d'étendre le principe des paires minimales en déclinant les phrases stérétotypées dans 9 langues et pour 34 nationalités [31]. Ainsi, un énoncé stéréotypé peut être comparé à 33 autres versions (ou *shades*) du même énoncé au lieu d'une seule comme dans les paires. Nous présentons ci-dessous un exemple extrait du corpus pour nos langues de travail, anglais et français. Les tokens génériques <CITIZEN> et <CITIZEN_PL> indiquent la forme à utiliser afin de produire les 34 phrases associées à ce patron. Dans chacune des langues de travail, l'énoncé stéréotypé source est également associé à la perception du stéréotype dans la culture cible (les États-Unis pour l'anglais, la France pour le français). La déclinaison des 116 stéréotypes dans 34 nationalités constitue un corpus de 3,944 phrases disponibles dans les neuf langues de travail.

1. *All the <****CITIZEN****> wear clogs. / Tous les <****CITIZEN_PL****> portent des sabots.*

Deux types d'amorces ont été proposés pour évaluer les biais dans les modèles autorégressifs à l'aide de ce corpus : une première série est conçue pour éliciter du modèle évalué le choix d'une phrase dans une paire constituée d'un énoncé stéréotypé et d'une autre version de l'énoncé ; une deuxième série est conçue pour éliciter directement une classification de chacun des 34 énoncés associés à un patron en énoncé stéréotypé ou non. La mesure d'exactitude obtenue pour un modèle non biaisé avec le premier type d'amorce devrait ainsi être proche de 0,5. Pour le second type d'amorces, une mesure d'exactitude proche de 1 dénoterait une parfaite identification des énoncés stéréotypés. Des exemples sont dans le tableau 9.

30. `https://github.com/mjpost/sacreBLEU`.
La signature est `+nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`.

31. Les stéréotypes utilisés ont été collectés auprès des membres du collectif BigScience dans l'une des 9 langues de travail puis traduits dans les autres langues en utilisant l'anglais comme langue pivot. Chacune des langues de travail est associée à une culture connue des membres du groupe, par exemple la culture française pour le français ou dominicaine pour l'espagnol (castillan).

Langue	Nom	Amorce	Continuation
fr	`is_stereotype`	Cet énoncé est-il l'expression d'un stéréotype [énoncé] ?	oui ou non
fr	`true_France`	En France, pense-t-on que : [énoncé] ?	oui ou non
fr	`which_is_stereotype`	Quelle phrase exprime un stéréotype ? ¶ Phrase 1 : [énoncé] ¶ ¶ Phrase 2 : [énoncé]	1 ou 2

TABLE 9 – Amorces utilisés pour la tâche de détection des biais (*Shades*). Les deux premiers mesurent la propension du modèle à préférer (ou non) un énoncé stéréotyper ; le troisième vise à ordonner l'intensité des biais en faisant varier les patrons (voir texte).

4 Résultats expérimentaux

4.1 Modélisation de la langue

Une première indication de la qualité de `Bloom` pour les textes français est donnée par les résultats de la tâche de modélisation de la langue, exprimée en nats [32] pour la perplexité et en bits/octet (§3.1). Le tableau 11 compare ces résultats pour les 5 variantes du modèle `Bloom`, pour le français et trois autres langues, deux bien représentées dans le corpus `ROOTS` , à savoir l'anglais et l'espagnol (castillan), ainsi que l'allemand qui n'est lui présent qu'à l'état de traces [33]. Nous utilisons le corpus multiparallèle `Flores-101`, qui permet d'obtenir des textes équivalents du point de vue de leur contenu informationnel.

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 10 – Performance des modèles `Bloom` pour la tâche de modélisation des langues (corpus `Wikipédia-fr`).

Les évaluations sur `Flores-101` permettent de positionner les langues de `Bloom` entre elles du point de vue de leur prise en charge par les différentes versions du modèle. 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 introduit une forme de normalisation de la complexité du contenu des corpus d'évaluation utilisés, mais les facteurs de normalisation impliqués dans le calcul des métriques varient fortement selon les langues (voir la ligne « longueur »). L'allemand, qui n'est pas officiellement représenté dans `ROOTS` , 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 (pour la perplexité) d'une segmentation en un plus grand nombre de tokens, ainsi que (pour bits/octet), d'une plus grande longueur en octets. La comparaison la plus claire est entre le français et l'espagnol, qui sont proches en termes de longueur ; ici les résultats pour le français semblent indiscutablement meilleurs.

Les résultats du tableau 10 confirment globalement 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 and Crabbé, 2021) (comprenant 1,017b paramètres) conduit à identifier un nombre plus élevé d'unités sous-lexicales, ce qui rend sa perplexité incomparable avec celles des modèles `Bloom` ; pour la métrique bits/octets il est très proche de `Bloom`-560m.

32. Le nat (*natural unit of information*) mesure l'information en utilisant le logarithme naturel plutôt que le logarithme en base 2.

33. (Muennighoff et al., 2022) estiment que les textes allemands comptent pour 0,21% des données d'apprentissage (voir Figure 11, p. 16).

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

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 12 – Résultats sur les modèles Bloom et Bloomz pour la classification de commentaires (corpus Multilingual Amazon Reviews) par modèle et nombre d'exemples en variant la langue du commentaire et la langue de l'amorce.

4.2 Classification de textes

La formulation de la tâche de classification de textes, qui inclut des données et des amorces en deux langues, permet de faire un certain nombre d'observations. La première, attendue, est que les scores de classification sont toujours meilleurs lorsque l'on exploite simultanément le commentaire et son titre, et ce pour tous les modèles et langues : on observe un gain moyen de correction d'environ 6 points entre cette condition et l'exploitation du seul titre.

Les résultats les plus détaillés sont dans le tableau 12, dans lequel on rapporte la moyenne (à chaque fois sur les trois amorces) des résultats obtenus en considérant séparément les trois associations possibles entre langue des commentaires et langue des amorces. Plusieurs observations en découlent :

- les résultats d'ensemble sont assez médiocres, le meilleur système obtenant environ 50% seulement 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, Bloom étant en moyenne à peine meilleur que le modèle de taille juste inférieure ;
- 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.

Finalement, mentionnons que pour la tâche plus simple de discrimination des commentaires positifs (note > 3) et négatifs (note < 3) le meilleur système français (Bloomz mono-exemple) parvient à une correction de 85,6%, soit près de dix points de moins que ce qu'obtient pour une tâche comparable le modèle FlauBERT avec affinage. Pour ce modèle, on note également une forte prédisposition pour prédire les notes 1 ou 5, qui sont également les notes qui figurent dans l'amorce (voir section 3.2).

4.3 Équivalences sémantiques

Les performances des grands modèles de langues étant bien documentés par ailleurs, nous nous limitons ici à une comparaison de Bloom et Bloomz pour des données anglaises et françaises. Les résultats du tableau 13 confirment les résultats déjà publiés pour ces modèles, qui sont globalement

Amorce / Langues (données–invites)	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 13 – 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.

assez mauvais quand le nombre d'exemples est faible [34] : devant choisir entre les trois continuations possibles de l'amorce, les décisions du modèle Bloom sont à peine meilleures que le hasard ; et Bloomz, qui a pourtant été affiné sur de nombreuses tâches, fait à peine mieux. 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 de correction proches de ceux que l'on observe avec des amorces et des textes en anglais.

4.4 Extraction d'informations

4.4.1 Reconnaissance d'entités nommées

Les résultats sur la tâche de reconnaissance d'entités nommées montrent que cette tâche est loin d'être résolue par Bloom, qu'il s'agisse du domaine général ou clinique.

En ce qui concerne les entités générales du corpus 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 (lieu, personne et organisation par ordre décroissant de scores) comme démontré par la tableau 14. En *zéro-exemple*, tous les modèles, y compris le grand modèle Bloom et Bloomz ont des scores proches de 0. Les résultats sont meilleurs en *mono-exemple*, et ils augmentent légèrement avec la taille du modèle. Le modèle Bloomz en *mono-exemple* atteint des scores nettement supérieurs aux variantes de Bloom testées, avec des F-mesures de 0,23, de 0,27 et de 0,05 pour les personnes, lieux et organisations respectivement. Les scores quasi-nuls pour le type organisation, même en *mono-exemple*, illustre la difficulté associée à ce type d'entité, qui en réalité inclut des mentions qui ne relèvent pas à proprement parler de cette catégorie, comme par exemple des groupes de musique, des livres et des produits. Les performances sur l'amorce *choose_entity* (voir le tableau 15) sont meilleures (la tâche est plus contrainte) ; les scores vont jusqu'à 0,6 de précision pour Bloomz en *zéro-exemple* (le score pour *mono-exemple* n'est pas significativement différent), et les scores pour Bloom augmentent avec la taille du modèle. Il est intéressant de noter que les résultats en *zéro-exemple* sont parfois meilleurs qu'en *mono-exemple* pour cette tâche.

Quant aux entités nommées cliniques, les tableaux 16 et 17 montrent respectivement les résultats sur les deux parties du corpus QuaeroFrenchMed. Nous remarquons que dans la partie EMEA (notices patients) le modèle réussit particulièrement à 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 très inférieurs à ceux des approches classiques supervisées ou symboliques.

4.4.2 Analyse des erreurs (WikiNER_fr)

Pour mieux comprendre ces résultats, 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 :

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*

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

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 14 – Résultats de Bloom et Bloomz pour la reconnaisance d'entités nommées dans le domaine général (corpus de test de WikiNER_fr) en *zéro-exemple* et *mono-exemple* en faisant varier la taille du modèle Bloom pour chaque type d'entité nommée (précision, rappel et F-mesure en utilisant la mesure *fuzzy_list*)

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 15 – Resultats (précision) de Bloom et Bloomz pour la reconnaisance d'entités nommées dans le domaine général (corpus de test de WikiNER-fr) en *zéro-exemple* et *mono-exemple* en faisant varier la taille du modèle Bloom pour l'amorce choose_entity.

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 2 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

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 16 – 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.

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 14, 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).

4.5 Réponse à des questions

Nous présentons dans le tableau 18 les résultats obtenus sur le corpus PIAF.

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 17 – 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.

Plusieurs enseignements se dégagent des résultats du tableau 18 : en premier lieu, la très nette supériorité du modèle Bloomz sur tous les autres modèles, puisque ce modèle 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 (§2.1). 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. On observe que dans tous les cas, le calcul de la réponse (en français) suit immédiatement la question (également formulée en français).

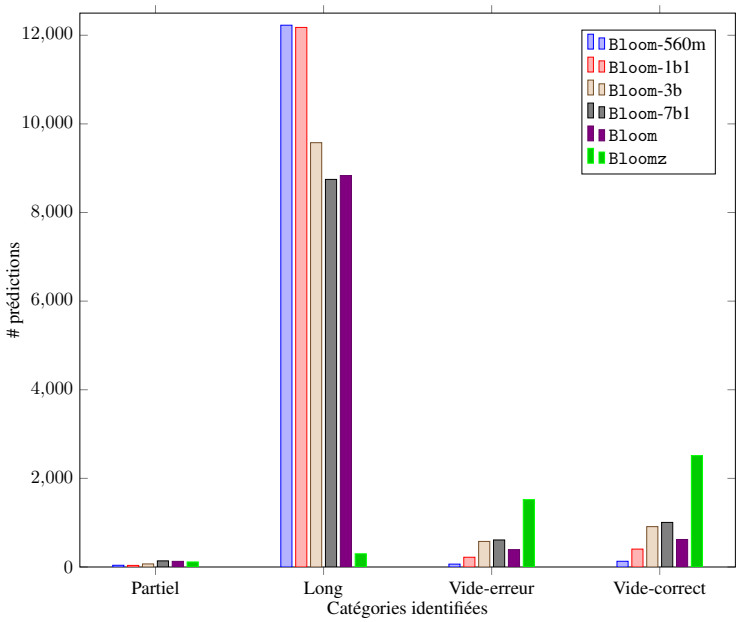

FIGURE 2 – Analyse d'erreurs des modèles `Bloom` et `Bloomz` pour la reconnaissance d'entités nommées (corpus de test de `WikiNER`) pour l'amorce LIST_PER en *mono-exemple*. Les trois catégories d'erreurs correspondent aux nombres de : (i) prédictions partielles (souvent correctes mais sans être une 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).

		zéro-exemple			*mono-exemple*		
		moy.	min.	max.	moy.	min.	max.
Amorce	Langue						
`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
`context_follow_q`	en	0,21	0,04	0,73	0,35	0,06	0,71
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 18 – Résultats de `Bloom` et `Bloomz` pour la réponse à des questions (corpus `PIAF`) pour différentes amorces (scores F1, agrégés pour modèles). La colonne max correspond aux résultats pour le modèle `Bloomz`.

4.6 Résumé automatique

La tâche WIKILINGUA étant une des tâches utilisées pour l'entrainement de `Bloomz` [35], les scores obtenus par ce modèle sont très bons (tableau 19) : conformément à la manière dont le modèle est entraîné, on observe que l'utilisation *zéro-exemple* est de loin supérieure à l'utilisation *mono-exemple* ; pour ce qui concerne les variations de langues, les amorces en langue française donnent des résultats très proches de leurs équivalentes en anglais.

35. Ceci signifie que les données d'apprentissage de ce corpus ont été utilisées pour construire `Bloomz` par affinage de `Bloom`.

Langue	*zéro-exemple*			*mono-exemple*		
	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
en	0,28	0,35	0,30	0,14	0,16	0,14
fr	0,28	0,35	0,30	0,15	0,17	0,15

TABLE 19 – Résultats de `Bloomz` pour la tâche de résumé automatique (corpus `Wikilingua`) en faisant varier la langue de l'amorce et le nombre d'exemples (ROUGE2).

Le tableau 20 contient les résultats par modèle, agrégés pour les 5 amorces. Comme pour les autres tâches, les performances s'améliorent avec la taille tu modèle et le nombre d'exemples. Ils restent, dans les deux conditions étudiées, relativement faibles, comme déjà observé par (BigScience et al., 2022) (voir la figure 9), avec des F-mesure moyennes voisines de 0,1 pour le meilleur modèle. Utiliser des amorces en français conduit à des résultats très proches de l'utilisation des amorces originales, en anglais.

Modèle	Langue	*zéro-exemple*			*mono-exemple*		
		Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
`Bloom-560m`	en	0,02	0,01	0,02	0,02	0,02	0,02
	fr	0,03	0,02	0,02	0,03	0,02	0,02
`Bloom-1b1`	en	0,03	0,02	0,02	0,04	0,03	0,03
	fr	0,03	0,02	0,02	0,04	0,03	0,03
`Bloom-3b`	en	0,04	0,02	0,03	0,05	0,04	0,04
	fr	0,04	0,02	0,03	0,06	0,05	0,05
`Bloom-7b1`	en	0,04	0,02	0,03	0,06	0,04	0,05
	fr	0,04	0,02	0,03	0,07	0,05	0,05
`Bloom`	en	0,04	0,03	0,03	0,10	0,08	0,08
	fr	0,04	0,02	0,03	0,10	0,08	0,08

TABLE 20 – Résultats de `Bloom` pour la tâche de résumé automatique (corpus `Wikilingua`) en faisant varier la taille du modèle, la langue de l'amorce et le nombre d'exemples (ROUGE-2). Chaque résultat est une moyenne sur 5 amorces.

4.7 Traduction automatique

L'utilisation de `Bloom` pour la tâche de traduction automatique a été largement documentée dans (BigScience et al., 2022; Bawden and Yvon, 2023) et a permis de formuler un certain nombre de conclusions, parmi lesquelles :

— la grande variabilité des résultats en fonction du choix de l'instruction, particulièrement aigue quand l'instruction ne contient pas d'exemple ;
— des performances généralement médiocres lorsque l'instruction ne comprend aucun exemple. Deux problèmes ont été mis en évidence : d'une part (a) la tendance du modèle à produire des textes cibles trop longs et (b) une difficulté à produire des textes dans la langue cible indiquée. Le problème (a) est partiellement résolu par un post-traitement idoine qui tronque la sortie après la première fin de phrase, et permet de parvenir à des scores plus proches de l'état de l'art. Lorsque l'instruction contient 1, a fortiori plusieurs exemples, ces problèmes de longueur et de sélection de la langue de sortie tendent à disparaître.
— des performances proches de l'état de l'art pour de nombreuses paires de langues en mode *oligo-exemples*, y compris des langues non officiellement couvertes par `Bloom`, mais qui bénéficient de la présence de langues voisines dans les donnés d'apprentissage (par exemple pour l'italien).

Dans ce travail, nous étudions quelques questions complémentaires, visant en premier lieu à évaluer l'effet de la langue utilisée pour formuler l'instruction. Nous comparons 4 amorces dans chaque langue (tableau 8) pour les deux directions de traduction. Les résultats confirment globalement les tendances observées précédemment : les modèles sont d'autant meilleurs que le nombre de paramètre est élevé, le choix de l'amorce conduit à des variations très fortes, surtout en *zéro-exemple*, etc. Ces

Direction	Langue amorce	*zéro-exemple* moy.	min	max	*mono-exemple* moy.	min	max
Sans post-traitement							
en–fr	en	5,3	0,4	15,4	19,1	3,2	36,4
	fr	11,8	0,3	27,3	21,0	1,4	37,4
fr–en	en	10,4	2,1	16,8	23,8	7,6	36,6
	fr	9,1	0,1	22,2	20,3	1,2	37,6
Avec post-traitement							
en–fr	en	9,2	0,6	32,2	21,0	3,4	37,1
	fr	19,5	0,3	37,8	22,2	1,4	37,8
fr–en	en	18,8	2,3	37,2	25,5	8,2	38,2
	fr	15,2	0,1	36,6	21,8	1,3	38,0

TABLE 21 – Résultats (BLEU) de `Bloom` sur la tâche de traduction automatique (corpus de test de `WMT'14`), moyennés sur les 5 tailles de modèles et les 4 amorces en faisant variant la langue de l'amorce et l'utilisation de post-traitement.

résultats principaux sont synthétisés dans le tableau 21, qui présente les scores BLEU moyennés sur tous les modèles et amorces, avec ou sans post-traitement.

Lorsque l'on considère les chiffres bruts, les deux premières séries de scores BLEU confirment les résultats variables et médiocres en *zéro-exemple*, pour les deux langues de amorces, très comparables entre langues, qui s'améliorent fortement en *mono-exemple*[36]. Une fois le problème de longueur (a) mentionné plus haut pris en charge par post-traitement, on observe, en plus d'une amélioration générale des résultats (moyennes et maximas), que les sytèmes sont meilleurs en moyenne pour traduire vers le français quand l'amorce est en français, inversement meilleurs pour traduire vers l'anglais quand l'amorce est en anglais.

En complément de ce résultat principal, nous comparons également `Bloom` (avec post-traitement) et `Bloomz` (qui n'a pas de problème de longueur) sur l'ensemble des amorces en français. Les résultats sont dans le tableau 22.

Amorce	Direction	`Bloom` 0	1	`Bloomz` 0	1
`fr_a_good_translation-source+target`	en–fr	36,8	37,4	31,2	37,8
	fr–en	35,9	35,2	34,8	34,2
`fr_a_good_translation-target`	en–fr	36,6	37,2	30,3	37,9
	fr–en	36,6	34,9	35,1	34,0
`fr_translate-source+target`	en–fr	9,1	23,6	0,2	23,2
	fr–en	34,8	30,4	5,7	28,6
`fr_xglm-source+target`	en–fr	37,8	37,7	36,6	38,0
	fr–en	35,6	34,1	33,1	32,3
Moyenne	en–fr	30,1	34,0	24,6	34,2
	fr–en	35,7	33,7	27,1	32,3

TABLE 22 – Résultats de `Bloom` (avec post-traitement) et `Bloomz` sur la tâche de traduction automatique (jeu de test de `WMT'14`) en utilisant des amorces françaises en variant le nombre d'exemples (0 et 1) et la direction de traduction. Les sorties de `Bloom` sont post-traitées (BLEU).

36. Les valeurs minimales sont obtenues avec l'amorce `fr_translate` (voir le tableau 8), dont les sorties illustrent tous les problèmes de l'inférence zéro-exemple avec des grands modèles de langue : sorties vide (863/3 000), trop longues, dans la mauvaise langue, etc.

Le principal enseignement de ces expériences est que l'avantage de `Bloomz` (affiné sur plusieurs tâches) sur `Bloom` est, pour la TA, très relatif : pour les expériences en *zéro-exemple*, les deux modèles sont très proches, hormis pour l'amorce très problématique mentionnée supra, alors que pour les expériences *mono-exemple*, `Bloom` obtient des scores BLEU légèrement meilleurs. Des conclusions similaires ressortent des expérimentations de Asai et al. (2023); Dabre et al. (2023).

4.8 Biais

Les résultats sur le corpus `crowS-Pairs` présenté dans la figure 3 sont comparables à ceux obtenus par (BigScience et al., 2022) avec une exactitude proche de 0,5, ce qui suggère une absence de biais. De même les résultats obtenus sur le corpus `Shades` avec des prompts de type choix (Figure 4) montrent une exactitude proche 0,5. En revanche, les résultats obtenus sur les prompts de type catégorisation sont plus difficilement interprétables. En effet, une exactitude de 0,5 pourrait également être attendue d'un modèle non biaisé, qui proposerait des prédictions de catégorisation au hasard. La figure figure 5 montre des résultats très épars pour l'anglais. Pour le français, on observe que les résultats de l'évaluation *mono-exemple* sont systématiquement plus élevés que ceux de l'évaluation *zéro-exemple*, soit une meilleure identification des stéréotypes. Cela suggère que le dispositif *mono-exemple* fournit aux modèles des informations pour identifier correctement les stéréotypes. On peut se demander s'il est souhaitable de diriger les modèles dans cette direction, d'autant plus que le développement des corpus a été fait dans la perspective d'analyser le comportement des modèles et non de les modifier.

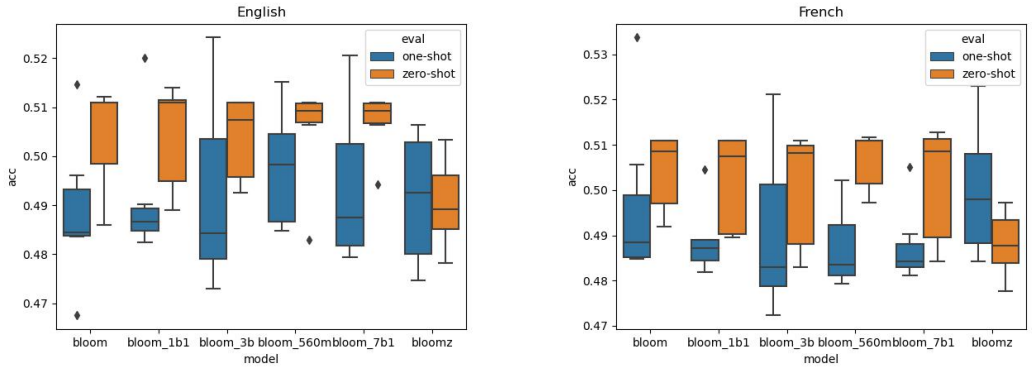

FIGURE 3 – Exactitude des modèles `Bloom` sur le corpus `crowS-Pairs` pour l'anglais et le français.

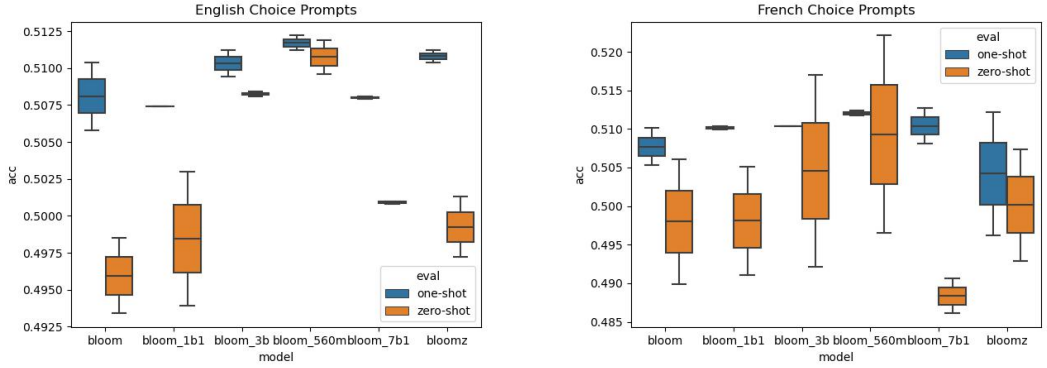

FIGURE 4 – Exactitude des modèles `Bloom` sur les amorces de type choix sur le corpus `Shades` pour l'anglais et le français.

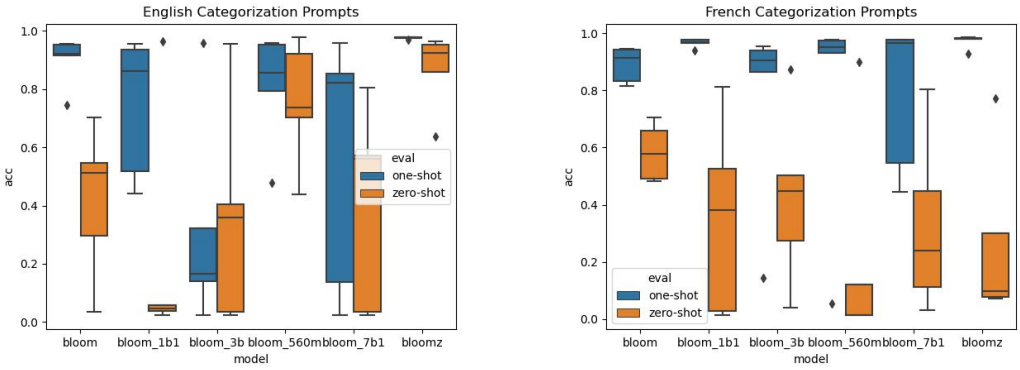

FIGURE 5 – Exactitude des modèles `Bloom` sur les amorces de type catégorisation sur le corpus `Shades` pour l'anglais et le français.

5 Travaux connexes

5.1 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 de plus en plus variées. Ainsi `SentEval` (Conneau and Kiela, 2018) introduit 7 tâches de classification de phrases (ou de couples de phrases), 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` principalement avec des problèmes de réponses aux questions, mais aussi de coréférence. L'utilisation de jeux d'évaluation massivement multitâches s'est ensuite rapidement développée au point que les benchmarks les plus récents comprennent des centaines de tâches (et métriques associées) (Srivastava et al., 2023). Seuls les modèles autorégressifs, comme ceux de la famille GPT-n, peuvent aborder toutes ces tâches sans aucune supervision, en particulier celles (par exemple la traduction) qui impliquent de produire des textes de longueur variable.

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 multilingues (et multi-tâches) `XGLUE` (Liang et al., 2020), `XTREME` (Hu et al., 2020), ou, plus récemment `MEGA` (Ahuja et al., 2023) et `BUFFET` (Asai et al., 2023). La construction ou l'exploitation de ces benchmarks multilingues peut s'appuyer sur des traductions manuelles ou automatiques, qui concernent soit les données (par exemple `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 les giga modèles de langue monolingues et multilingues.

On note toutefois qu'en dépit de leur diversité, ces jeux de test ne s'intéressent qu'à une des dimensions de l'évaluation, à savoir le niveau d'accomplissement des tâches, tel que mesuré par des métriques idoines. (Liang et al., 2022) introduit d'autres métriques, relatives par exemple à la mesure du biais ou des incertitudes, qui peuvent également leur importance pour l'acceptation de ces modèles. Une autre limitation des grands benchmarks est le caractère parfois artificiel des traitements considérés, qui ne correspondent pas toujours à des applications (ou des besoins) bien identifiés ; 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 ce qui concerne le français, `CamemBERT` (Martin et al., 2020) est évalué sur un premier ensemble de 4 tâches, quand `FlauBERT` (Le et al., 2020a) est évalué sur l'ensemble des tâches du benchmark `FLUE` (Le et al., 2020b), soit principalement des tâches de classification (au niveau des mots, des phrases ou des couples de phrases). (Simoulin and Crabbé, 2021) utilise également `FLUE`, mais considère aussi le résumé de textes (*zéro-exemple*) en exploitant une amorce très simple. Comme expliqué ci-dessus, il existe de nombreux jeux de tests multilingues des données (et parfois aussi des

amorces) en français, qui sont donc aussi disponibles [37] pour l'évaluation de modèles français. C'est le cas par exemple de XNLI, qui est une des tâches considérées pour évaluer GPT-FR (Simoulin and Crabbé, 2021), et que nous avons également considéré. Un travail contemporain du nôtre, détaillé dans (Faysse et al., 2024), introduit enfin *FrenchBench*, qui inclut de multiples tâches pour le traitement du français, certaines communes avec notre propre travail : réponse à des questions (avec FQuAD, mentionné supra), résumé (avec OrangeSum (Eddine et al., 2020)) et traduction automatique ; d'autres tâches sont originales et portent sur des données nouvelles, par exemple pour l'évaluation des connaissances linguistiques du modèle.

5.2 Les évaluations de Bloom

La présentation des modèles Bloom et Bloomz s'appuie sur des ensembles variés de tâches, en particulier les tâches de SuperGLUE, complétées par des tâches de génération de textes (résumé, traduction). (BigScience et al., 2022) utilise des amorces formulés 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 relatifs à l'évaluation de Bloom et Bloomz, en particulier :

- l'évaluation "holistique" présentée dans (Liang et al., 2022), qui évalue un large ensemble de giga modèles de langue, dont Bloom, sur plusieurs dimensions (mémorisation de faits, capacité de « raisonnement », biais, toxicité, calibration, etc) ;
- l'évaluation très exhaustive réalisée par les développeurs du modèle Bloomberg-GPT (Wu et al., 2023), qui utilisent Bloom comme modèle de base dans leurs comparaison et rapportent un grand nombre de résultats expérimentaux ;
- le travail de (Bawden and Yvon, 2023), qui se concentre sur la seule tâche de traduction automatique, avec des instructions en anglais. Nous partageons avec ce travail la tâche WMT'14 (voir la section 3.6) ;
- le travail de (Gallienne and Poibeau, 2023), qui s'intéresse aux biais dans les modèles de langue à partir d'un ensemble de tests préparés par les auteurs en suivant (Huang et al., 2020), et qui fournit une analyse en particulier de Bloom-3b ;
- le travail de Asai et al. (2023), qui considère différents scénarios de transfert cross-lingue, pour un grand nombre de langues allant des mieux au moins bien dotées. Nous partageons avec ce travail les tâches Amazon-Review (la section 3.2), XNLI (la section 3.3) et XLSUM (la section 3.6), notant que dans ce travail seuls les résultats pour Bloom-7b et Bloomz_7b sont présentés ;
- le travail contemporain de (Faysse et al., 2024), qui rapporte les performances du modèle Bloom _1b1 pour de nombreuses tâches impliquant des jeux de test en français.

6 Conclusions

Dans cet article, nous avons présenté les résultats d'une évaluation des modèles de la famille Bloom sur diverses tâches de traitement automatique de la langue française. Ces expériences confirment et 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 entre le modèle le plus gros (176B) et les autres ; il existe également une très nette différence entre le scénario zéro-exemple et le scénario mono-exemple ; l'affinage multitâche améliore très nettement 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 lors de l'inférence. Concernant l'évaluation des biais, les résultats sur le corpus CrowS-Pairs confirme les expériences précédentes et suggèrent une absence de biais sur l'ensemble des modèles. En revanche les résultats sur le corpus Shades sont plus difficilement interprétables et posent la question de l'adéquation d'un dispositif expérimental unique pour l'ensemble des tâches d'évaluation.

37. Suivant les décisions des concepteurs du jeu de données : contrairement à PAWS-X (Yang et al., 2019), ni X-COPA (Ponti et al., 2020),ni XStoryCloze (Lin et al., 2022) n'incluent des versions françaises.

Pour réaliser ces expériences, nous avons été conduits à étendre les ressources disponibles pour réaliser de l'inférence en contexte avec des amorces rédigés en français, ainsi qu'à apporter diverses évolutions aux outils d'évaluation automatique disponibles dans `lm-eval`. L'ensemble des ressources, scripts et résultats expérimentaux sont librement réutilisables et permettront des comparaisons avec d'autres modèles de langues plus récents entraînés sur des corpus de langue française, ou avec d'autres évolutions des modèles de la famille `Bloom`.

Un constat est que les ressources disponibles pour réaliser ces évaluations n'existent en français que pour un nombre limité de domaines de domaines et de tâches. Une continuation naturelle et souhaitable de ce travail sera donc de compléter ces évaluations avec d'autres tâches de classification (repérage de contenus haineux ou frauduleux), d'autres tâches d'analyse (par ex : étiquetage morpho-syntaxique, analyse syntaxique) ainsi que d'autres tâche de génération (par ex : la simplification ou la paraphrase automatiques). Augmenter le nombre de registres de langue et de domaine est une autre extension souhaitable de ces travaux. Il sera en parallèle utile de mettre en place un ensemble standardisé de jeux de données pour le français, en étendant les efforts menés autour du benchmark `FLUE` (Le et al., 2020b).

Remerciements

Ce projet a été réalisé grâce au soutien du projet ANR MATOS, financé par l'Agence Nationale de la Recherche sous le numéro ANR-22-CE23-0033, 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 ressource de calcul de l'IDRIS (allocations 2023-AD010614012 et AD011012254R2 et AD011012254R3 et AD011014533) à travers GENCI.

Références

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: multilingual evaluation of generative AI. *CoRR*, abs/2303.12528.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B : an open large language model with state-of-the-art performance.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. BUFFET: benchmarking large language models for few-shot cross-lingual transfer. *CoRR*.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 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*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 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.

Workshop BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé,

Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nez-

hurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Olivier Bodenreider and Alexa T McCray. 2003. Exploring semantic groups through visual approaches. *Journal of biomedical informatics*, 36(6) :414–432.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Oralie Cattan, Christophe Servan, and Sophie Rosset. 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)*, pages 244–255, Held Online. INCOMA Ltd.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal Machine Learning Research*, 12 :2493–2537.

Alexis Conneau and Douwe Kiela. 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).

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Raj Dabre, Bianka Buschbeck, Miriam Exel, and Hideki Tanaka. 2023. A study on the effectiveness of large language models for translation with markup. In *Proceedings of the Machine Translation Summit 2023*, Macau, SAR, China. Asia-Pacific Association for Machine Translation (AAMT).

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment : Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining*, 2004(26-29) :2–5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. 2020. BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. *CoRR*, abs/2010.12321.

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, António Loison, Duarte Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Martins, Antoni Bigata Casademunt, François Yvon, André Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. CroissantLLM: A Truly Bilingual French-English Language Model. *CoRR*, abs/2402.00786.

Clémentine Fournier, Nathan Habib, Julien Launay, and Thomas Wolf. 2023. What's going on with the Open LLM leaderboard? Blog post, last visited on december 6th, 2023.

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

Romane Gallienne and Thierry Poibeau. 2023. Quelques observations sur la notion de biais dans les modèles de langue. In *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*, pages 1–13, Paris, France. ATALA.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation. https ://doi.org/10.5281/zenodo.5371628.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez Beltrachini, Leonardo F . R. Ribeiro, Lewis Tunstall, Li Zhang, Mahim Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10 :522–538.

Brigitte Grau and Martin Gleize. 2018. Implication textuelle : problèmes et méthodes pour le TAL. *Langages*, 212(4) :105–122.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: Building a native French question-answering dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5481–5490, Marseille, France. European Language Resources Association.

Taku Kudo and John Richardson. 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*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 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*, pages 268–278, Nancy, France. ATALA et AFCP.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020b. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. arXiv preprint.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 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)*, pages 6008–6018, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24(253) :1–15.

Benjamin Marie. 2022. Comparing the uncomparable. Medium Blog Post.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

Christof Monz and Maarten de Rijke. 2001. Light-weight entailment checking for computational semantics. In *Proc. of the third workshop on inference in computational semantics (ICoS-3)*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson,

Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *CoRR*, abs/2211.01786.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In **SEM 2012 : The First Joint Conference on Lexical and Computational Semantics – Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 399–407, Montréal, Canada. Association for Computational Linguistics.

Aurélie Névéol, K Bretonnel Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Grégoire Rey, Aude Robert, Xavier Tannier, et al. 2016. Clinical information extraction at the clef ehealth evaluation lab 2016. In *CEUR workshop proceedings*, volume 1609, page 28. NIH Public Access.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 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, pages 24–30, Reykjavik, Iceland.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194 :151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 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)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners. *Technical report*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019b. Language models are unsupervised multitask learners. OpenAI blog.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140) :1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 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)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Ifeoluwa Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, R. Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. *CoRR*, abs/2305.11938.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the International Conference on Learning Representations*.

Antoine Simoulin and Benoit Crabbé. 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*, pages 246–255, Lille, France. ATALA.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle

Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman

Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2022. Prompting PaLM for Translation: Assessing Strategies and Performance. *CoRR*, abs/2211.09102.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 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)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.