FLOR: On the Effectiveness of Language Adaptation

Severino Da Dalt[1*], Joan Llop[1*], Irene Baucells[1*], Marc Pàmies[1*], Yishi Xu[2], Aitor Gonzalez-Agirre[1], Marta Villegas[1]
[1]Barcelona Supercomputing Center
[2]Cerebras Systems

Abstract

Large language models have amply proven their great capabilities, both in downstream tasks and real-life settings. However, low- and mid-resource languages do not have access to the necessary means to train such models from scratch, and often have to rely on multilingual models despite being underrepresented in the training data. For the particular case of the Catalan language, we prove that continued pre-training with vocabulary adaptation is a better alternative to take the most out of already pre-trained models, even if these have not seen any Catalan data during their pre-training phase. We curate a 26B tokens corpus and use it to further pre-train BLOOM, giving rise to the FLOR models. We perform an extensive evaluation to assess the effectiveness of our method, obtaining consistent gains across Catalan and Spanish tasks. The models, training data, and evaluation framework are made freely available under permissive licenses.

Keywords: LLMs, Language Adaptation, Vocabulary Adaptation, Continued Pre-Training, Catalan, Spanish

1. Introduction

Over the past few years, transformer-based language models have dominated a wide range of Natural Language Processing (NLP) tasks. Their widespread use has now expanded beyond the NLP community, increasing the number of user interactions in all sorts of languages. However, most of these models have been developed exclusively for English, with much less efforts being devoted to other languages. Non-English speakers often have to rely on multilingual models that were trained on a mix of languages, as their more desirable monolingual counterparts may not exist due to the prohibitive amount of language-specific data and compute power required to generate them.

Despite the cross-lingual transfer capabilities of multilingual models, adding too many languages to the mixture can be detrimental for downstream performance (Conneau et al., 2019). Moreover, models that cover hundreds of languages (e.g. XLM (Lample and Conneau, 2019), mT5 (Xue et al., 2020) or PaLM (Chowdhery et al., 2022)) require bigger vocabulary sizes, typically five times larger than single-language vocabularies. This results in bigger embedding layers, which in the case of relatively small models represent a higher percentage of the total number of parameters and, thus, the memory footprint of the model can be significantly increased.

In the context of a mid-resource language like Catalan, we study the most prominent possibilities: *from scratch* and *continued pre-training*, both with and without vocabulary adaptation. By changing the vocabulary, the resulting model benefits from language-specific tokenization at the cost of reusing fewer embeddings from the original model. We believe that this trade-off is worth exploring, given that recycling model weights is arguably preferable to randomly initializing them.

We then report performances on a set of standard NLP tasks to test the capabilities of the models built throughout our experiments, shedding some light on the differences between language adaptation strategies. The results show that, for our particular linguistic context, vocabulary adaptation is the optimal strategy to efficiently recycle a publicly available model.

Overall, the main contributions of this work are:

- FLOR-760M[1] and FLOR-1.3B[2], two autoregressive language models that achieve state-of-the-art results in several Catalan and Spanish downstream tasks, when compared to open models of similar size.
- A curated corpus with 26 billion tokens of Catalan, Spanish and English text[3].
- A novel evaluation benchmark for Catalan and Spanish decoder-only models.
- A series of experiments that lead us to build the FLOR models, namely: 1) comparing the from scratch initialization strategy against language adaptation, 2) studying the two main language adaptation approaches, and 3) determining if a monolingual (Cerebras-GPT) or multilingual (BLOOM) model is more suitable for language adaptation.

*Equal contribution.

[1]huggingface.co/projecte-aina/FLOR-760M
[2]huggingface.co/projecte-aina/FLOR-1.3B
[3]huggingface.co/datasets/BSC-LT/open_data_26B_tokens_balanced_es_ca

LREC-COLING 2024, pages 7377–7388
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

All the aforementioned resources, including those developed for experimental purposes, are openly released under permissive licenses for commercial usage and further research.

2. Related Work

The most widespread strategy to pre-train large language models is to start from scratch with randomly initialized weights. Examples of this can be found across many languages, such as Spanish (Gutiérrez-Fandiño et al., 2022), Russian (Emelyanov et al., 2020), French (Launay et al., 2022), Arabic (Antoun et al., 2021) or Chinese (Zeng et al., 2021).

However, in low-resource scenarios, this tendency to start training from scratch seems unjustified. Continued pre-training presents itself as a viable alternative to reuse existing models by extending their next-word-prediction training with no architectural or tokenizer changes. The literature offers a few examples of successfully adapting an English model to a different language. Notably, Müller and Laurent (2022) demonstrated this by tailoring GPT-J for French, while Pires et al. (2023) showcased a similar success in their work on adapting LLaMA for Portuguese.

Multilingual models trained from scratch are also very present in the NLP landscape (Shliazhko et al., 2023; Lin et al., 2022a; Scao et al., 2022). They can achieve a high level of cross-lingual transfer but tend to require much larger vocabularies and can suffer from some limitations such as the curse of multilingualism (Conneau et al., 2019) or an unfair distribution of languages (Choudhury and Deshpande, 2021).

In this multilingual context, Yong et al. (2023) adapts BLOOM to 8 additional languages. The authors compare continued pre-training with two adapter techniques, namely MAD-X (Pfeiffer et al., 2020) and $(IA)^3$ (Liu et al., 2022). Their results show that adapter-based language adaptation strategies are preferable for models that exceed the 3B parameters, but no conclusions are reached for smaller sizes. Ebrahimi and Kann (2021) conducted similar experiments with encoder models of smaller size and concluded that continued pre-training is more promising than both vocabulary extension and MAD-X.

On the other hand, Artetxe et al. (2020) show that deep monolingual models can generalize across languages without relying on joint training or a shared vocabulary. In a similar direction, de Vries and Nissim (2021) study three language adaptation methods based on retraining lexical embeddings from a monolingual model before further pre-training on new languages. Later research has expanded upon this line, with Minixhofer et al. (2022) introducing a method that employs semantic similarity to pair embeddings from the source and target languages.

Lakew et al. (2018) proposed a simple, yet effective strategy for adapting the embedding layer to a new tokenizer by reusing the embeddings corresponding to the shared tokens. Ostendorff and Rehm (2023) go one step further initializing the non-matching tokens as the weighted average of the shared tokens' embeddings, using a fully-trained smaller model as a reference. The weight of each token is approximated using the similarity of token embeddings in the smaller model. This comes with the non-trivial cost of requiring a fully trained smaller model.

In summary, existing vocabulary transfer techniques aim at reusing the weights of an existing model and adapting its tokenizer to a new language. This not only intends to boost performance but also provides enhanced inference speed, reduces the word-to-token ratio and can decrease the total size in the case of multilingual models (Gee et al., 2022).

3. Data

3.1. Pre-training Corpus

The mid-resource nature of the Catalan language hinders the collection of sufficient data and motivates us to add Spanish text to the pre-training corpus, given their cultural closeness and high linguistic similarity. In this way, we also ensure that the resulting corpus reflects the bilingualism present in the Catalan society. In particular, our custom-made dataset has 42.1% of the total data in Catalan and 41.3% in Spanish. The remaining 16.6% is English text, which is used as an anchor between the source and target tokenizers. The addition of a third language also intends to prevent catastrophic forgetting of a language that is already mastered by the original model. As it can be seen in Table 3.1, the corpus contains a mixture of several data sources in an attempt to increase domain diversity. A more detailed description of each source can be found in Appendix A.

Overall, the training corpus contains roughly 26B tokens. Note that the choice of corpus size is not arbitrary at all, since it is the required amount of tokens to train a Chinchilla-optimal 1.3B model, according to the scaling laws proposed by Hoffmann et al. (2022). In order to reach this amount of tokens with a balanced Catalan-Spanish ratio, a slight oversampling of the Catalan data was required. However, no more than 4 epochs were given to any of the data sources, respecting the rule-of-thumb given by Muennighoff et al. (2023).

It is relevant to note that the checkpoints that are used as a starting point had already seen billions of tokens in their respective pre-trainings. For reference, Cerebras-GPT was trained on 26.3B tokens

Dataset	Lang.	Epochs	Tokens (M)
Wikipedia	CA	3.5	1127.08
C4_ca	CA	2.1	8381.54
Biomedical	CA	1.4	23.33
VilaWeb	CA	2.1	149.29
CaWaC	CA	2.1	171.41
Racó - Notícies	CA	2.1	50.89
Racó - Fòrums	CA	2.1	989.81
Wikipedia	ES	1.4	1371.43
C4_es	ES	0.1	7805.53
Biomedical	ES	0.7	449.85
Legal	ES	0.7	984.37
Gutenberg	ES	0.7	52.57
Wikipedia	EN	1.4	4290.56
Total			25847.66

Table 1: Data sources in the training corpus with their respective number of tokens.

of English texts from The Pile dataset (Gao et al., 2020), and BLOOM on 341B multilingual tokens from the ROOTS corpus (Laurençon et al., 2022). For experimental purposes, we also assembled a 2B tokens dataset made exclusively of Wikipedia text, with a perfect balance between Catalan, Spanish and English.

3.2. Preprocessing details

The 26B tokens of curated text were obtained after applying a series of data-cleaning steps to ensure compliance with quality standards. The first step was to perform language filtering using fastText's linear classifier (Joulin et al., 2016). Then all corrupted Unicode characters were normalized with the ftfy library (Speer, 2019). Deduplication is then performed at the document level with Onion (Pomikálek, 2011). Finally, a series of filters were applied in order to discard poor-quality data (e.g. very short documents or paragraphs, sentences with undesired characters, etc). As a final step, potential pornographic content from the scrapped sources was filtered out by a tailor-made RoBERTa classifier (Liu et al., 2019). The resulting corpus is publicly released under an open license.

3.3. Evaluation

We perform an extensive evaluation to compare ourselves against other publicly available models of similar size. To do so, we rely on the open-source codebase released by EleutherAI (Gao et al., 2021), and extend their evaluation framework with 10 additional datasets. Many of these are recently published Catalan datasets (Gonzalez-Aguirre et al., 2024) that have not yet been used for evaluation, and the rest are publicly available third-party datasets that were not present in EleutherAI's framework at the time of writing. In all cases, we assess model performance with commonly used metrics in a 5-shot setting.

The resulting benchmark is a collection of standard downstream tasks in Catalan, Spanish and English, enabling a fair comparison across languages. In particular, it includes tasks for reading comprehension, commonsense reasoning, question answering, natural language inference, paraphrase identification and machine translation. Although for some discriminative tasks, such as NLI, QA, and Paraphrase Identification, decoder models are not the most suitable option, we add these tasks to our benchmark to better understand the capabilities of the model. For more details about the evaluation datasets, refer to Appendix B. We openly release the evaluation scripts on GitHub[4] for future use.

4. Methodology

This section describes the training procedure and presents the experiments that were carried out to prove the effectiveness of the chosen method.

First, Section 4.1 describes the base models used for continued pre-training. In Section 4.2, we analyze the efficiency of our custom tokenizer. And, finally, in the remainder of this section, we describe our three pre-train strategies: from scratch (Section 4.3), continued pre-training (Section 4.4) and vocabulary adaptation (Section 4.5).

4.1. Models

For the main experimental setup we use the Cerebras-GPT architecture (Dey et al., 2023), which is basically a GPT-3-like model that employs dense attention in all decoder blocks, rather than alternating dense and sparse-banded attention like the original GPT-3 (Brown et al., 2020). The publicly released checkpoint has been pre-trained on 26.3B tokens of English text, which makes it chinchilla-optimal (Hoffmann et al., 2022). We selected this model for all language adaptation experiments because it has never seen Catalan or Spanish data during its pre-training phase. Thus, it can be used as a proof-of-concept to show that the language adaptation techniques presented in this work can also be applied to unseen languages.

In pursuit of producing the best Catalan model in the one billion parameter range, we apply the best-performant language adaptation strategy to the BLOOM-1.1B and BLOOM-1.7B models (Scao et al., 2022), as we expect this strategy can benefit from their multilingual capabilities. These decoder-only models use AliBi positional embeddings and layer normalization after the embedding layer. They have been pre-trained on 341B tokens

[4] https://github.com/projecte-aina/flor_language_adaptation

of data in 46 natural languages and 13 programming languages. This corpus is estimated to contain 36.83B tokens of Spanish data and 3.75B of Catalan.
In both cases, the tokenizers were trained with the byte-level Byte Pair Encoding (BPE) algorithm (Sennrich et al., 2016). Cerebras' models reuse the GPT-2 tokenizer, which has a vocabulary size of 50,257, whereas BLOOM models have a vocabulary size of 250,680 tokens, a common practice in highly multilingual models to reduce the risk of word over-segmentation, especially if they contain many languages with different scripts.

4.2. Tokenization

We train a new tokenizer running byte-level BPE on the same corpus that we later use for pre-training (see Section 3.1). We set the vocabulary size to 50,257 in all cases.
For Cerebras-GPT-1.3B, Table 2 shows that replacing the tokenizer greatly reduces the average number of tokens-per-word (TPW) in Catalan (∼30%) and Spanish (∼33%). The increase experienced in English is much smaller (∼16%), an affordable cost considering that English is only used as an anchor and is not among our target languages.

Tokenizer	ca	es	en
Cerebras_50k	2.19	2.13	**1.41**
Catalan_50k	**1.53**	**1.43**	1.64
BLOOM_250k	1.42	1.37	*1.37*
Catalan_250k[5]	**1.39**	**1.28**	*1.37*

Table 2: Average Tokens-Per-Word of the source and target tokenizers in each language.

In the case of BLOOM, on the other hand, we have a slight TPW increase both in Catalan (∼8%) and Spanish (∼4%). This is compensated by the fact that, as a consequence of the original BLOOM vocabulary being roughly five times larger, we achieve a model size reduction of approximately 29% and 27% for BLOOM-1.1B and BLOOM-1.7B, respectively. Table 3 depicts the level of compression achieved in each case.
As a final note, it is relevant to highlight that our new tokenizer shares 26.35% of its vocabulary with the one from Cerebras-GPT-1.3B, and 66.16% with BLOOM's. This leads us to believe that we are facing successful adaptation, and expect that to be reflected in the final results.

4.3. From Scratch

We test the most widely used strategy by training a GPT-like model from scratch. More specifically, a Cerebras-GPT model of 1.3 billion parameters is trained on 26B tokens of data, complying with the well-known Chinchilla scaling laws.

[5]Only shown for comparison, not actually used.

Model	**V_{size}**	**Total Size**
BLOOM-1.1B	250k	1.07B
FLOR-vocab_adapted-750M	50k	0.76B (-29%)
BLOOM-1.7B	250k	1.72B
FLOR-vocab_adapted-1.3B	50k	1.31B (-24%)
Cerebras-GPT-1.3B	50k	1.32B
Cerebras-GPT-vocab_adapted-1.3B	50k	1.32B (±0%)

Table 3: Model sizes before and after vocabulary adaptation.

4.4. Continued Pre-training

Continued pre-training is the most straightforward and widely used approach for language adaptation, and numerous such examples can be found in the literature (Pires et al., 2023; Müller and Laurent, 2022; la Rosa and Fernández). It involves extending the pre-training phase of a LM with data in a new language while preserving the original weights and vocabulary. However, a significant drawback is that it requires the use of a tokenizer that was not originally designed for the target language.
Given the monolingual nature of Cerebras-GPT, its tokenizer performs poorly when given Catalan or Spanish data. The higher TPW ratio significantly reduces the amount of text that can fit into the model's input sequence, which in turn slows down training and increases inference costs. In light of this, when training Cerebras-GPT from scratch, the full pre-training dataset is encoded in 38B tokens, as opposed to the 26B tokens we get from our custom tokenizer.

4.5. Vocabulary Adaptation

For this strategy, we reuse the weights of the source model corresponding to the transformer layers and we reinitialize the embedding layer, for which a number of studies have explored effective strategies. A core element in some of the most prominent proposals aims to avoid random initialization of token embeddings (Minixhofer et al., 2022; Ostendorff and Rehm, 2023; Gee et al., 2022; Lakew et al., 2018). Given a source vocabulary and a target vocabulary, we copy the weights of the embeddings corresponding to all tokens present in both vocabularies and initialize the rest to the average of all source embeddings weights.
In order to fully integrate the newly adapted embeddings into the model and avoid instabilities during pre-training, specific training strategies are proposed in the literature. de Vries and Nissim

(2021) trains only the embedding weights while freezing the rest of the model layers, while Howard and Ruder (2018) slowly unfreezes the rest of the layers during training. We make a preliminary experiment, testing the efficiency of each of these strategies. In this setup, we adapt the Cerebras-GPT-1.3B and train it using the subset of Catalan, Spanish, and English Wikipedias, mentioned in Section 3.1. A full description of the strategies follows:

Only embeddings. We further train the embedding layer while freezing the transformer layers. This is the approach used in Artetxe et al. (2020), with the difference that they randomly initialize the lexical embeddings; according to the authors, freezing the Transformer layers helps to avoid catastrophic forgetting.

Vocabulary adaptation. Following the line of Ostendorff and Rehm (2023), which emphasizes that freezing most parameters limits the model's ability to learn about the new language, we directly train all model weights on the Wikipedia subset.

Progressive unfreezing. An intermediate approach between the previous ones, consisting of training the model while progressively unfreezing the model layers.

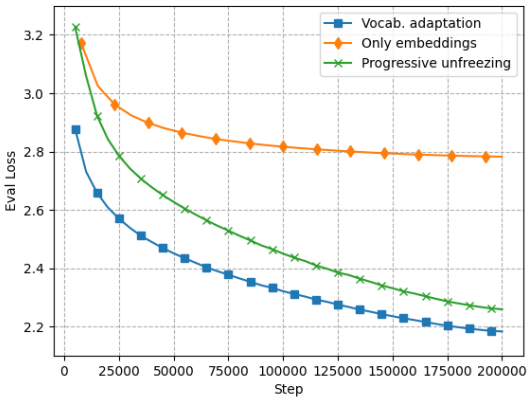

Figure 1: Evaluation loss across our three vocabulary adaptation strategies.

In Figure 1, we compare the three strategies on the basis of the validation loss during training. The results suggest that freezing the layers significantly reduces training speed, while no training instabilities appear. It is possible that, for smaller models, altering the embedding layer implies changing a substantial proportion of the model, which might potentially lead to training instabilities. However, we have not observed any such occurrences. Notably, the vocabulary adaptation strategy achieves the lowest validation loss.

5. Results

Table 4 presents the evaluation results obtained on the benchmark introduced in Section 3, categorized by task type. Several state-of-the-art models of similar size are included as strong baselines, as well as a random baseline for reference. Cerebras-GPT-1.3B (Dey et al., 2023) and BLOOM-1.1B (Scao et al., 2022) are added for being the source models used in our language adaptation strategies, as detailed in Section 4. Additionally, we evaluate four English models, namely GPT-Neo-1.3B (Black et al., 2022), Pythia-1.4B (Biderman et al., 2023), OPT-1.3B (Zhang et al., 2022), and Falcon-rw-1.3B (Penedo et al., 2023), along with the multilingual model mGPT-1.3B (Shliazhko et al., 2023). Proprietary models were purposely excluded from the comparison.

The results show that the FLOR family of models, starting with the 760M model and followed by the 1.3B model, consistently achieves the highest results in the Catalan tasks. The only exceptions are PAWS-X, where all scores are very close to random, and XNLI, where the BLOOM-1.1B base model performs slightly better FLOR-760M. In the Spanish-to-Catalan translation task, FLOR models also lag slightly behind the vocabulary-adapted version of Cerebras-GPT. FLOR models dominate on most Spanish tasks, but this superiority does not extrapolate to the English domain, where, unsurprisingly, the English-only models consistently outperform the others.

When comparing the results of the Cerebras-GPT model trained from scratch with the language-adapted models, the former outperform in all tasks and languages except for paraphrase identification tasks. In the case of language-adapted models, a comparison between continued pre-training and vocabulary adaptation strategies reveals that vocabulary-adapted models consistently outperform their counterparts, with the sole exceptions being the English Belebele and PAWS-X datasets. In addition to the evaluations on downstream tasks, we monitor the evolution of the perplexity metric during training to compare strategies across our three languages of interest (see Figure 2). The vocabulary-adapted models achieve lower perplexities a lot faster than their from-scratch counterparts. However, according to previous work (Yong et al., 2023), perplexity during language adaptation training may not consistently align with prompting performance, for which our conclusions regarding these procedures do not hinge solely on this metric.

6. Discussion

Choosing an effective initialization strategy. The two language adaptation methodologies tested in this work, i.e. continued pre-training

Model	Training tokens		Reading Comprehension			Question Answering				
			Belebele$_{acc}$			XQuAD$_{f1}$			CatalanQA$_{f1}$	CoQCat$_{f1}$
	Pre-train	Lang. adapt.	ca	es	en	ca	es	en	ca	ca
Random	-	-	25.00	25.00	25.00	-	-	-	-	-
mGPT-1.3B	440B	-	26.11	24.44	26.11	0.33	0.67	0.17	0.65	0.78
GPT-Neo-1.3B	380B	-	35.44	32.77	41.67	19.75	29.77	51.53	22.34	23.57
Pythia-1.4B	299.9B	-	35.78	35.11	41.56	26.19	34.13	52.98	27.47	25.38
OPT-1.3B	180B	-	35.22	33.56	43.78	23.53	31.85	52.95	26.58	20.18
Falcon-rw-1.3B	350B	-	34.33	33.56	**47.89**	5.93	19.25	**58.60**	6.91	15.61
Cerebras-GPT-1.3B	26B	-	33.44	31.89	36.67	8.56	19.98	36.00	10.87	14.12
BLOOM-1.1B	341B	-	39.89	37.22	39.33	36.81	36.98	44.10	44.65	34.57
From_scratch-1.3B	26B	-	33.44	31.00	29.00	8.93	8.47	4.19	13.58	18.86
Cerebras-GPT-continued_pre-training-1.3B	26B	38B	39.22	35.11	33.56	26.67	28.17	25.22	34.99	31.93
Cerebras-GPT-vocab_adapted-1.3B	26B	26B	40.33	36.22	35.33	28.52	30.38	28.27	39.99	39.50
FLOR-760M	341B	26B	41.00	37.89	37.00	41.10	41.11	40.20	51.01	41.34
FLOR-1.3B	341B	26B	**43.44**	**39.11**	40.44	**43.52**	**44.31**	44.11	**54.25**	**48.15**

Model	Training tokens		Natural Language Inference				Paraphrase Identification			
			XNLI$_{acc}$			TE-ca$_{acc}$	PAWS-X$_{acc}$			Parafraseja$_{acc}$
	Pre-train	Lang. adapt.	ca	es	en	ca	ca	es	en	ca
Random	-	-	33.33	33.33	33.33	33.33	50.00	50.00	50.00	50.00
mGPT-1.3B	440B	-	40.06	43.81	45.67	37.03	51.00	52.30	56.15	51.32
GPT-Neo-1.3B	380B	-	41.44	45.57	49.92	35.38	54.65	53.40	54.60	51.70
Pythia-1.4B	299.9B	-	42.46	45.61	51.00	37.46	54.15	52.50	**57.70**	55.23
OPT-1.3B	180B	-	40.08	44.53	**52.48**	36.14	54.10	52.55	55.90	53.23
Falcon-rw-1.3B	350B	-	34.53	35.85	45.73	34.96	54.25	**54.05**	53.65	50.60
Cerebras-GPT-1.3B	26B	-	36.83	38.88	47.25	35.62	52.40	52.20	55.95	52.05
BLOOM-1.1B	341B	-	47.19	46.39	49.44	41.38	**55.05**	54.05	54.75	55.65
From_scratch-1.3B	26B	-	43.77	42.24	38.40	38.78	51.45	51.35	53.55	54.15
Cerebras-GPT-continued_pre-training-1.3B	26B	38B	45.55	44.01	41.48	40.53	53.50	51.40	50.35	53.95
Cerebras-GPT-vocab_adapted-1.3B	26B	26B	46.21	45.61	43.35	42.65	49.95	50.85	51.25	56.30
FLOR-760M	341B	26B	46.93	46.03	46.11	42.14	52.35	52.50	54.85	56.55
FLOR-1.3B	341B	26B	**49.20**	**48.82**	47.45	**42.89**	53.20	52.85	53.00	**57.43**

Model	Training tokens		Commonsense Reasoning				Translation					
			XStoryCloze$_{acc}$		COPA$_{acc}$		FLoRes$_{bleu}$					
	Pre-train	Lang. adapt.	es	en	ca	en	ca→es	es→ca	ca→en	en→ca	es→en	en→es
Random	-	-	50.00	50.00	50.00	50.00	-	-	-	-	-	-
mGPT-1.3B	440B	-	55.33	60.09	52.20	63.40	3.25	2.96	9.25	3.79	17.75	15.34
GPT-Neo-1.3B	380B	-	51.42	66.58	53.40	74.80	3.27	3.80	17.77	5.49	17.70	12.04
Pythia-1.4B	299.9B	-	54.14	68.37	52.20	78.60	9.68	5.74	24.03	11.10	21.50	15.04
OPT-1.3B	180B	-	53.94	69.95	52.60	76.20	3.14	3.52	15.39	2.00	16.33	6.53
Falcon-rw-1.3B	350B	-	51.09	**71.34**	52.40	**79.60**	3.03	3.59	8.89	3.01	14.17	6.50
Cerebras-GPT-1.3B	26B	-	49.11	60.62	51.40	66.80	2.42	1.81	2.69	0.82	3.36	1.77
BLOOM-1.1B	341B	-	57.91	62.48	62.80	66.40	21.62	15.28	31.16	21.28	20.92	16.84
From_scratch-1.3B	26B	-	55.20	53.54	61.40	59.60	2.72	2.22	1.36	1.14	1.51	1.08
Cerebras-GPT-continued_pre-training-1.3B	26B	38B	56.45	57.38	61.60	60.80	9.19	14.88	18.23	12.14	13.10	7.71
Cerebras-GPT-vocab_adapted-1.3B	26B	26B	58.64	59.10	66.40	61.60	16.31	**19.63**	26.65	24.10	17.16	15.09
FLOR-760M	341B	26B	61.42	61.42	65.40	64.20	**22.62**	15.77	32.26	26.04	20.91	18.08
FLOR-1.3B	341B	26B	**64.06**	61.81	**68.00**	67.80	22.16	18.58	**33.95**	**29.31**	**23.09**	**20.30**

Table 4: Evaluation results on Catalan, Spanish and English downstream tasks in a 5-shot setting. The upper part of each table contains the baseline models, including the random one whenever possible. The bottom part of each table displays the in-house models. The underline indicates the training tokens from our pre-training corpus explained in Section 3. The *Cerebras-GPT-continued_pretraining-1.3B** corresponds to the checkpoint trained on the same number of tokens as the other strategies but on fewer data, as it has a different tokenizer.

and vocabulary adaptation, involve reusing transformer layer weights from a pre-trained model. Both demonstrated superior performance compared to our model trained from scratch, which has a random initialization. This is yet another example in which random initialization of weights proves not to be the most efficient strategy.

The importance of language-specific tokenization.

We compare both language adaptation strategies: with and without vocabulary adaptation. As explained in Section 4.4, the FLOPs needed to see

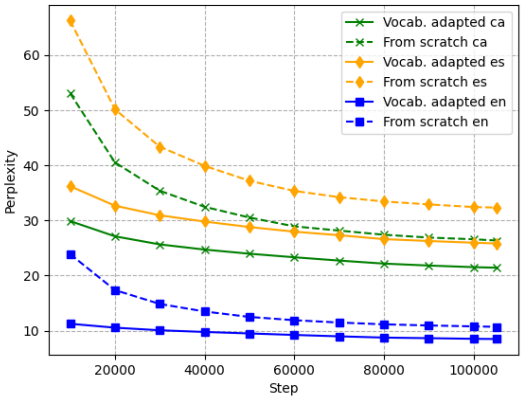

Figure 2: Evaluation perplexity comparison of our vocabulary adaptation and from scratch models across the three main languages.

the same amount of data can increase due to not having a language-specific tokenizer. This is probably behind the fact that the vocabulary-adapted model outperforms the continued pre-trained model trained using the same FLOPs in most tasks and languages. What's more interesting, the vocabulary-adapted model also outperforms the continued pre-trained model trained with the full dataset. We find it surprising that English tasks also benefit from vocabulary adaptation to the Catalan tokenizer. Considering that both models originate from the same checkpoint and have been exposed to identical data, the sole remaining distinction lies in the tokenizer, with one tokenizer adapted to the language and the other not. Consequently, we add yet another piece of evidence (Artetxe et al., 2020) indicating that the impact of tokenization on downstream task performances is significant.

Leveraging a multilingual model for language adaptation. Vocabulary adapted models built from BLOOM outperformed those built from Cerebras-CPT model. This is expected mainly because of the larger pre-training corpus seen by BLOOM, which already includes Catalan and Spanish data and related languages (several other Romance languages). An additional factor that may be favoring FLOR models is a higher token overlap between the source and target tokenizers in comparison to Cerebras-CPT, which enables more lexical embeddings to be reused. Moreover, it must be taken into account that this vocabulary adaptation, departing from such large multilingual tokenizer, results in a substantial reduction in target lexical embeddings, which translates into a significant model distillation. Specifically, when adapting the 1.3B BLOOM model to our vocabulary, it downscaled to a 760M-parameter model, while the 1.7B BLOOM model was reduced to 1.3B parameters.

On preserving English in language adapted models. The downstream tasks results in Table 4, show that language-adapted models tend to decrease their performance on English tasks compared to their source model, most likely due to the Catalan and Spanish-specific training during language adaptation. Although keeping some English data in training may help to mitigate the performance drop, it remains uncertain without further experiments. For the Cerebras-based models, the negative impact of vocabulary adaptation on performance varies between tasks, with some showing small declines (e.g., XStoryCloze and Belebele) and others significant drops (e.g., XQuAD and XNLI). The reasons for these differences are unclear. Similarly, the FLOR-760M models, except for PAWS-X, perform worse in English than the BLOOM-1.1B models. On the other hand, unexpectedly, when vocabulary adaptation and continued pre-training are compared on the same data, the latter tends to perform worse in English. In translation tasks, which can be considered an exception due to the bilingual nature of the task, the adaptation of the Cerebras-GPT-based models allows a radical improvement in results, essentially transforming them from incompetence to satisfactory results. In the case of BLOOM models, which possess baseline translation capabilities, vocabulary adaptation further enhances their performance.

The paraphrase identification tasks on the evaluation benchmark repetitively fall to random. The difficulty of this task was demonstrated by the fact that, in general, even the best models often struggled to beat the random baseline. This is particularly true for the multilingual PAWS-X. We also acknowledge that, ideally, paraphrasing should be evaluated in its generative form. However, it is well-recognized that automatically assessing generative tasks, including summarization, presents considerable challenges and requires further research (Scao et al., 2022).

7. Conclusions

In this work, we pursue the best strategy to develop a 1.3B generative language model in Catalan using a 26B token Catalan, Spanish, and English corpus, released under an open license. We use this data to train a model from scratch and compare it to adapting an existing pre-trained generative LM to our target language. The results demonstrate the superiority of the latter option, proving a significant improvement over a random initialization of weights.

Within language adaptation, we then compare two main methodologies: *continued pre-training*,

where the target-language corpus is used to further train the source model with its original weights and tokenizer, and *vocabulary adaptation*, where the transformer layer weights are kept but the embeddings reinitialized to fit a new tokenizer created for the target language. Vocabulary adaptation proves to be the most effective technique in terms of performance gains and training efficiency, provided a more language-specific tokenization of the target data. Finally, we compare vocabulary adaptation using two source model architectures, Cerebras-GPT and BLOOM, monolingual and multilingual, respectively, and conclude the superiority of the latter, which leads to our best models: FLOR-760M and FLOR-1.3B. We believe that these experiments and findings, focused on Catalan, can be useful to other languages in order to guide the more efficient development of their own models.

For all our experiments, we developed an evaluation benchmark that focuses on Catalan, but also includes Spanish and English tasks. We cover different evaluation capabilities and leave it to future work to include more generative tasks with sensitive evaluation metrics. Furthermore, our efforts should prioritize the unresolved issues that have emerged from our discussions, namely the impact of including a small fraction of English data among the target training corpora and the explanation of why vocabulary adaptation improves over continued pre-training even in the source language.

8. Acknowledgments

This work has been promoted and financed by the Generalitat de Catalunya through the Aina Project.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335 y 2022/TL22/00215334.

This work is supported by DeepR3 (TED2021-130295B-C31), a project funded by the MCIN/AEI/10.13039/501100011033 and the European Union NextGeneration EU/PRTR program.

9. Bibliographical References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, and Marta Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12710–12718.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

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.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. 2023. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages.

Anton Emelyanov, Tatiana Shavrina, Oleh Shliazhko, and Artem Snegirev. 2020. Russian GPT-3 models. `https://github.com/ai-forever/ru-gpts`.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling.

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.

Leonidas Gee, Andrea Zugarini, Leonardo Rigutini, and Paolo Torroni. 2022. Fast vocabulary transfer for language model compression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 409–416, Abu Dhabi, UAE. Association for Computational Linguistics.

Aitor Gonzalez-Aguirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of catalan. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. European Language Resources Association and the International Committee on Computational Linguistics.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish legalese language model and corpora.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Javier De la Rosa and Andres Fernández. Zero-shot reading comprehension and reasoning for spanish with BERTIN GPT-J-6B. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Conference*

on Spoken Language Translation, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Julien Launay, E.l. Tommasone, Baptiste Pannier, François Boniface, Amélie Chatelain, Alessandro Cappelli, Iacopo Poli, and Djamé Seddah. 2022. PAGnol: An extra-large French generative model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4275–4284, Marseille, France. European Language Resources Association.

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, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

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. 2022a. 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.

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. 2022b. 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.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nikola Ljubesic and Antonio Toral. 2014. cawac - A web corpus of catalan and its application to language modeling and machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 1728–1732. European Language Resources Association (ELRA).

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models.

Martin Müller and Florian Laurent. 2022. Cedille: A large autoregressive french language model.

Malte Ostendorff and Georg Rehm. 2023. Efficient language model training through cross-lingual and progressive transfer learning.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.

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. *J. Mach. Learn. Res.*, 21(1).

Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mgpt: Few-shot learners go multilingual.

Robyn Speer. 2019. ftfy. Zenodo. Version 5.5.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer.

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.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

A. Data Sources

List of data sources from the training corpus:

- **Wikipedia**: English, Spanish and Catalan articles covering a wide range of topics. Lists, tables, hyperlinks and boilerplate were removed.
- **mC4**[6]: Catalan and Spanish portions of the multilingual Common Crawl's web crawl corpus (Raffel et al., 2020).
- **Biomedical data**: Spanish biomedical corpus crawled in 2020. The sources were medical journals, pharmaceutical companies, research centers and health-related websites, among others (Carrino et al., 2021).
- **Legal data**: Spanish texts from the legislative, advocacy and administrative domains, mostly scraped from digital resources. It includes documents such as patents, criminal proceedings or the Spanish Consitution (Gutiérrez-Fandiño et al., 2021).
- **Gutenberg**[7]: Spanish books with no copyright retrieved from the Gutenberg Project's digital library.
- **VilaWeb**[8]: News articles from a Catalan-language daily news outlet.
- **CaWaC**[9]: A 780M token web corpus of Catalan text built from the *.cat* top-level-domain in late 2013. (Ljubesic and Toral, 2014).
- **Racó Català**[10]: Posts from the News and Forum sections of a Catalan web portal.

B. Evaluation Datasets

List of datasets from the evaluation benchmark:

- **Belebele** (Bandarkar et al., 2023) is a reading comprehension dataset covering 122 language variants, including Spanish and Catalan. Each document consists of a passage of text from the FLORES-200 dataset, together with a question and four options for multiple-choice answers.
- **XNLI** (Conneau et al., 2018) is one of the biggest Natural Language Inference (NLI) corpus that spans across 15 languages. Recently, a professional Catalan translation was added to the corpus (Gonzalez-Aguirre et al., 2024).
- **COPA** (Roemmele et al., 2011) is a dataset designed to assess commonsense causal reasoning. Each document consists of a premise and two alternatives, and the goal is to determine which of them has a more plausible causal relationship with the premise. COPA has been professionally translated into Catalan to be incorporated into our evaluation benchmark (Gonzalez-Aguirre et al., 2024).
- **XStoryCloze** (Lin et al., 2022b) is the professional translation of the English StoryCloze dataset into 10 languages, including Spanish, and is used to assess commonsense reasoning with a particular focus on story comprehension. It consists of selecting the correct ending from among two options for a four-sentence story.
- **XQuAD** (Artetxe et al., 2020) is a multilingual extractive question answering dataset with parallel coverage in 11 languages, including Spanish. The benchmark has latterly been extended to Catalan via professional translation.
- **CoQCat** (Gonzalez-Aguirre et al., 2024) is a dataset for Conversational Question Answering in Catalan, just recently released. The task consists in answering a series of questions interconnected in a conversation about a text passage.
- **CatalanQA** (Gonzalez-Aguirre et al., 2024) is an extractive question answering dataset built in Catalan using text passages from news articles and the Catalan Wikipedia, each of which is associated with between 1 and 5 questions.
- **PAWS-X** (Yang et al., 2019) is a multilingual paraphrase identification dataset, consisting of sentence pairs categorized as paraphrases or non-paraphrases. It is available in six languages, including Spanish, and has recently been made available in Catalan through professional translation (Gonzalez-Aguirre et al., 2024).
- **Parafraseja** (Gonzalez-Aguirre et al., 2024) is a paraphrase identification dataset built in Catalan from sources originally written in that language.
- **FLoRes** (Team et al., 2022) is a machine translation dataset consisting of parallel sentences in multiple pairs of English and low-resource languages. For our evaluation, we used the Flores200 version of the dataset and restricted our focus to the six language combinations between English, Spanish and Catalan.

[6] https://huggingface.co/datasets/mc4
[7] https://gutenberg.org
[8] https://www.vilaweb.cat
[9] https://huggingface.co/datasets/cawac
[10] https://www.racocatala.cat