Probing the Emergence of Cross-lingual Alignment during LLM Training

Hetong Wang[ϵ] **Pasquale Minervini**[ϵ] **Edoardo M. Ponti**[ϵ,κ]
[ϵ]University of Edinburgh [κ]University of Cambridge
H.Wang-197@sms.ed.ac.uk

Abstract

Multilingual Large Language Models (LLMs) achieve remarkable levels of zero-shot cross-lingual transfer performance. We speculate that this is predicated on their ability to align languages without explicit supervision from parallel sentences. While representations of translationally equivalent sentences in different languages are known to be similar *after convergence*, however, it remains unclear how such cross-lingual alignment emerges *during pre-training* of LLMs. Our study leverages intrinsic probing techniques, which identify which subsets of neurons encode linguistic features, to correlate the degree of cross-lingual neuron overlap with the zero-shot cross-lingual transfer performance for a given model. In particular, we rely on checkpoints of BLOOM, a multilingual autoregressive LLM, across different training steps and model scales. We observe a high correlation between neuron overlap and downstream performance, which supports our hypothesis on the conditions leading to effective cross-lingual transfer. Interestingly, we also detect a degradation of both implicit alignment and multilingual abilities in certain phases of the pre-training process, providing new insights into the multilingual pretraining dynamics.[1]

1 Introduction

Language Models (LMs) pre-trained on unlabelled multilingual texts show remarkable performance in zero-shot cross-lingual transfer (BigScience Workshop et al., 2023; Xue et al., 2021; Conneau et al., 2020a). In fact, fine-tuning a multilingual LM on annotated data for a downstream task in a source language allows it to perform inference in other target languages, too—although often with varying degrees of degradation (Pires et al., 2019; Wu and Dredze, 2019; Libovický et al., 2019; Wu and Dredze, 2020). Surprisingly, this occurs even when the vocabularies of two languages have a null intersection, i.e., no tokens are shared (Artetxe et al., 2020). Similarly, if the model scale is sufficiently large, LLMs are able to perform cross-lingual transfer through in-context learning with few examples in the source language (Lin et al., 2022).

This implies that LMs can implicitly align lexica and grammar between languages even in the absence of explicit parallel data. To explain this ability, previous work showed that multilingual LMs can encode texts from different languages into language-agnostic representations (Muller et al., 2021, *inter alia*) and that grammatical functions are encoded in the same subsets of neurons (Stanczak et al., 2022). Nevertheless, the existing literature mainly examined the final model upon convergence. Thus, they fail to explain how cross-lingual alignment *emerges* during self-supervised pre-training and how this impacts zero-shot cross-lingual transfer in downstream tasks.

Hence, our study aims to explore the dynamics of cross-lingual alignment throughout pre-training, discovering trends such as those shown in Figure 1. First, we adopt a reliable intrinsic metric for cross-lingual alignment, namely the extent to which morphosyntactic features (e.g., *Number* for nouns or *Tense* for verbs) tend to activate the same subnetwork within LMs. This implies that the more two languages are aligned, the higher the overlap of the subsets of neurons encoding their information.

We speculate that the degree of alignment tends to increase during pre-training, and that this facilitates the emergence of zero-shot transfer capabilities. To corroborate this hypothesis, we calculate the correlation between intrinsic metrics of alignment and cross-lingual downstream task performance. To identify neuron overlap at different stages of pre-training, we rely on intrinsic probing (Stańczak et al., 2023). Specifically, we probe several checkpoints of BLOOM (BigScience Workshop et al., 2023), a prominent multilingual LM.

[1]Our code is available at: https://github.com/ErikaaWang/probing-multilingual-dynamics

Findings of the Association for Computational Linguistics ACL 2024, pages 12159–12173

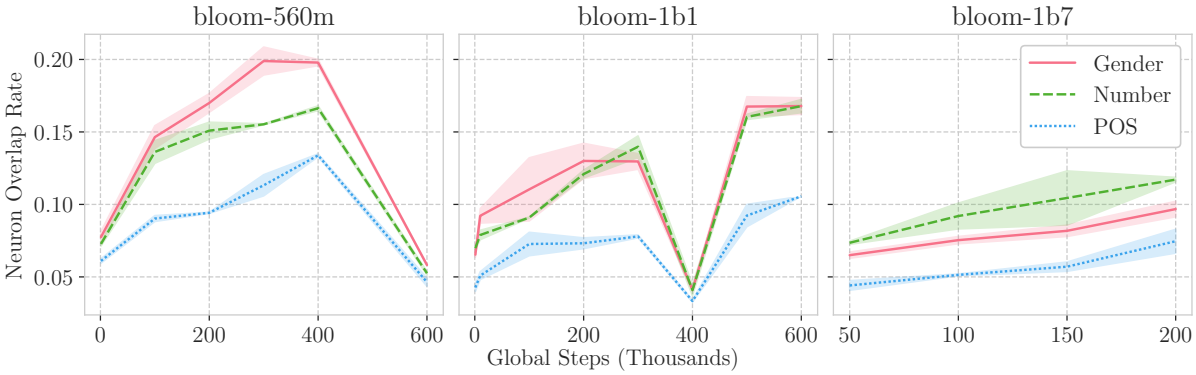


Figure 1: Neuron overlap rates (averaged across pairs of languages) across pre-training steps. Colours and line styles identify selected morphosyntactic categories. The three plots correspond to different model scales.

This also allows us to compare the emergence of cross-lingual alignment at different model scales, from small (560m) to medium-sized (1.7B) LMs.

To measure the relation between implicit alignment and downstream performance, we evaluate the zero-shot cross-lingual transfer ability of these checkpoint models on part-of-speech tagging in 11 languages from Universal Dependencies (UD; Nivre et al., 2017) and natural language inference in 7 languages (XNLI; Conneau et al., 2018). We find a statistically significant, strong correlation between neuron overlap and downstream performance across all model scales. Furthermore, we report a somewhat unexpected finding: both metrics do not grow monotonically during pre-training; rather, they may experience severe drops both in the middle and at the end of pre-training in smaller model scales.

2 Intrinsic Probing

We first aim to identify the subnetworks that each language activates within LLMs. To this end, we employ the latent variable model proposed by Torroba Hennigen et al. (2020) for intrinsic probing, which can identify the subset of specific dimensions within a representation that encodes the information for a particular linguistic feature. Formally, given a dataset $\mathcal{D} = \{(\pi^{(n)}, \boldsymbol{h}^{(n)})\}_{n=1}^{N}$, where $\boldsymbol{h}^{(n)} \in \mathbb{R}^d$ are d-dimensional embeddings and $\pi^{(n)} \in \Pi$ are labels that belong to an inventory for a particular linguistic feature (e.g., a part of speech or a morphosyntactic category), our goal is to probe representations $\boldsymbol{h}$, with a total neuron set of $D = \{1, \ldots, d\}$, to identify the subset $C^\star \subseteq D$ that contains the k most informative neurons with respect to the linguistic feature Π. For example, the labels $\Pi = \{\text{Singular}, \text{Plural}\}$ are associated with the morphosyntactic category of *Number*. In our setup, we extract hidden representations $\boldsymbol{h}$ from $\text{BLOOM}_{\text{560m}}$, $\text{BLOOM}_{\text{1b1}}$ and $\text{BLOOM}_{\text{1b7}}$. Thus, $d \in \{1024, 1536, 2048\}$, respectively.

Since we are interested in probing the subset of most informative neurons C, we introduce a latent variable $C \subseteq D$ in the probe $p_{\boldsymbol{\theta}}(\pi \mid \boldsymbol{h})$:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\pi \mid \boldsymbol{h}) &= \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi, C \mid \boldsymbol{h}) \\ &= \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi \mid \boldsymbol{h}, C)\, p(C), \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}$ are the parameters of the probe. Following the optimal settings in Stanczak et al. (2022), we choose a uniform distribution for the prior $p(C)$.

To estimate the parameters $\boldsymbol{\theta}$, directly optimising its log-likelihood in Equation (1) is intractable since it requires marginalising over all possible k-sized subsets C of D, which grow as $\binom{d}{k}$. Thus, we optimise its variational lower bound instead:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}\left(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}\right) \\ &\geq \sum_{n=1}^{N} \left(\mathop{\mathbb{E}}_{C \sim q_\phi} \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \right] + H\left(q_\phi\right) \right), \end{aligned} \tag{2}$$

where $H(q_\phi)$ is the entropy of q_ϕ, a variational distribution over C parameterised by ϕ.[2] Stańczak et al. (2023) showed that the Poisson sampling is a practically efficient sampling scheme for $q_\phi(C)$, in which each dimension is considered to be independently sampled from a Bernoulli distribution. Therefore, we opt for the Poisson sampling scheme in our setup.

[2]The full derivation is available in Stańczak et al. (2023).

After having trained the probe model $p_{\boldsymbol{\theta}}(\pi \mid \boldsymbol{h})$ on the morphosyntactic category Π, we determine the most informative subset $C^\star$ by maximising the posterior:

$$C^\star = \underset{C \subseteq D, |C|=k}{\arg\max} \sum_{n=1}^{N} \log p\left(\pi^{(n)} \mid \boldsymbol{h}_C^{(n)}\right) \quad (3)$$

where $\boldsymbol{h}_C$ is the masked sub-vector of $\boldsymbol{h}$ that contains only dimensions in C. Since the above combinatorial optimisation problem is intractable in practice, we use a greedy search method for selecting neurons 1 to k.

3 Experimental Setup

Models. We conduct the following experiments on BLOOM (BigScience Workshop et al., 2023), an open-access autoregressive multilingual LM that is jointly trained on data from 46 natural languages and 13 programming languages. The list of covered languages and their ISO codes is available in Appendix A.1. In particular, we consider three model sizes: 560m, 1b1, and 1b7, with 6, 8, and 4 valid intermediate model checkpoints, respectively.[3] Both the checkpoints of $\text{BLOOM}_{\text{560m}}$ and $\text{BLOOM}_{\text{1b1}}$ spread evenly from *1k* to *600k* global training steps, and $\text{BLOOM}_{\text{1b7}}$ ranges from *1k* to *300k*, where the global batch size is increased from 256 to 512. Moreover, BLOOM models with different sizes are trained on an equivalent amount of tokens—which is around 341 billion from the ROOTS corpus (Laurençon et al., 2022)—and share the same tokenizer. All these configuration designs allow us to consistently study their training trajectories across scales.

We studied the cross-lingual ability of BLOOM through two metrics: (i) neuron overlap between languages (Section 3.1); (ii) zero-shot cross-lingual transfer performance on XNLI and on POS tagging (Section 3.2), which require multilingual semantic and syntactic knowledge, respectively. In the next section, we will report how these metrics change across pre-training steps and how they correlate with each other.

3.1 Intrinsic Probing

Data. In order to collect the dataset $\mathcal{D}$ mentioned in Section 2, we take advantage of annotated sentences from Universal Dependencies (UD) treebanks v2.1 (Nivre et al., 2017) from 13 languages. The UD labels are first mapped to the UniMorph Schema (Kirov et al., 2018) by the converter proposed by McCarthy et al. (2018) to ensure a unified label scheme across languages. Then, we compute the contextual representations of each word by BLOOM at selected layers. If words are tokenised into subwords, we represent them as the average of their token embeddings, following Vulić et al. (2020). After that, these embedding–label pairs are grouped by linguistic feature (part of speech, number, gender, etc.) and randomly shuffled. Finally, they are split into train, validation, and test sets so that words with the same lemma (e.g. *eat* and *ate*, *employ* and *employer*) appear in the same set. This avoids trivial memorisation of lemma-related information during probe training. Additionally, words with lemmas occurring less than 20 times in a split are discarded. This procedure finally results in a batch of datasets $\mathcal{D}$, each corresponding to a particular morphosyntactic feature, a specific language, and a specific layer depth from which representations are extracted. The available language–feature pairs are listed in Appendix A.2.

Training. An individual probe is trained for each dataset $\mathcal{D}$ to identify the neurons that encode most information for the corresponding morphosyntactic feature in a specific language. The probes are trained on the training set with the objective function in Equation (2). The probe with parameters $\boldsymbol{\theta}$ is a linear projection followed by a softmax:

$$p\left(\pi^{(n)} \mid \boldsymbol{h}_C^{(n)}\right) = \text{softmax}\left(W \boldsymbol{h}_C^{(n)}\right) \quad (4)$$

where $W \in \mathbb{R}^{|\Pi| \times d}$. After training the probes, neuron sets $C^\star$ are chosen greedily on validation sets by Equation (3), where $\boldsymbol{h}_C^{(n)}$ is performed by masking all non-chosen dimensions as zero. Based on the results presented in Stanczak et al. (2022), we set $k = 50$ as a compromise between performance and computational efficiency.

Cross-lingual Alignment Metric. We use the average overlap rate over all possible language pairs for a specific morphosyntactic feature as the metric for cross-lingual alignment. Specifically, we compute the overlap rate of the 50 dimensions probed for each possible language pair where a morphosyntactic feature is expressed, as listed in Appendix A.2. Each category will result in an overlap rate

[3] https://huggingface.co/bigscience/bloom-intermediate. We discovered that the released checkpoints of 1) $\text{BLOOM}_{\text{560m}}$ at steps 10k and 500k, 2) $\text{BLOOM}_{\text{1b7}}$ at steps 1k and 10k, and 3) $\text{BLOOM}_{\text{1b7}}$ at steps 250k and 300k are duplicate model pairs. Thus, we remove these invalid models from the checkpoint collection to ensure reliability.

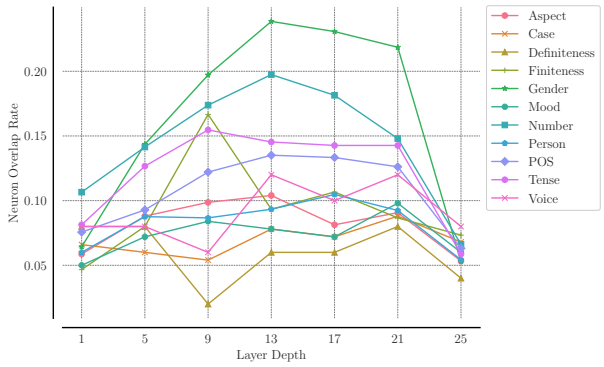


Figure 2: The extent of alignment through layers in the converged $\text{BLOOM}_{560\text{m}}$.

matrix. Examples are displayed as heatmaps in Appendix B.1.

Selection of Layers and Linguistic Features. To focus on selected layers and linguistic features, we first exhaustively examine 7 equally distributed layers of the converged $\text{BLOOM}_{560\text{m}}$ on 11 morphosyntactic features by intrinsic probing. Figure 2 illustrates the extent of cross-lingual alignment throughout different layer depths. By comparing the average pairwise overlap rate among linguistic features, the neurons that encode information about *Number* and *Gender* overlap the most, peaking at layers 13 and 17 out of 25. Other linguistic features, such as *Case*, *Mood* and *Tense* display a more even trend throughout layers, amounting to 7%, 8% and 15% on average. There are also fluctuations around layer 9, where the alignment of *Finiteness* jumps to a peak, while *Voice* and *Definiteness* decrease sharply.

Incidentally, we observe a drastic decrease in overlap rates at the last hidden layer across all 11 features. This phenomenon contrasts with the trend observed in encoder-only models such as m-BERT and XLM-R (Stańczak et al., 2023; Stanczak et al., 2022), where a significant overlap is observed at the output layer. This difference is intuitive, as it aligns with the training objective of different model architectures: encoder-only models are optimised on a Masked Language Modelling objective to replicate the original token, whereas autoregressive models, such as BLOOM, are trained on Next Token Prediction objectives.

Based on the aforementioned results, we select the features *Number* and *Gender*, as well as layers 13 and 17, for our experiments on checkpoints, as they are overall the most informative. Additionally, we include *POS* tags as a linguistic feature, as it provides the largest language coverage. The other two model scales (1b1 and 1b7) also have the same amount of total layers (25), which allows us to adopt an identical policy in layer selection.

3.2 Cross-lingual Transfer Evaluation

We evaluate the zero-shot cross-lingual transfer ability of the checkpoint collections of BLOOM by two kinds of downstream tasks. Similar to Hu et al. (2020), we focus on single-source transfer: the annotated training and validation data is provided only in the source language, English. The trained model is directly tested on target languages. We opt for (i) the XNLI dataset as a sentence classification task, and (ii) POS tagging as a structured prediction task. Both are part of widespread multilingual benchmarks such as XTREME (Hu et al., 2020): we follow the same data splits.

- **XNLI** The Cross-lingual Natural Language Inference dataset, dubbed XNLI (Conneau et al., 2018), is designed to evaluate the sentence understanding abilities in target languages by determining the relationship between two sentences. The relationships considered are whether the premise *entails*, *contradicts*, or is *neutral* towards the hypothesis.

- **POS** Part-of-speech tagging data is sourced from UD treebanks (Nivre et al., 2017). These treebanks consist of sentences in a wide range of languages, where each word is annotated with one of the 17 universal POS tags.

Training. We finetune each checkpoint with the same hyperparameter setting to ensure a fair comparison of their cross-lingual transfer ability. We use the AdamW (Loshchilov and Hutter, 2019) optimiser with a learning rate of 2×10^{-5}. The models are trained for 5 epochs on XNLI with a training set of 392k samples and 10 epochs on POS tagging with a training set of 21k samples. We perform model selection based on development set performance, evaluating models every 100 (POS) or 500 (XNLI) steps. Finally, we evaluate the best finetuned models on the test set of each target language, using accuracy as a metric for XNLI and F1 score for POS tagging.

Limited by computational resources, we conduct this fine-tuning with qLoRA (Dettmers et al., 2023). This first quantises the (frozen) pre-trained LLM to 4-bit and then applies a trainable Low-Rank Adapter (LoRA; Hu et al., 2022). The adapter

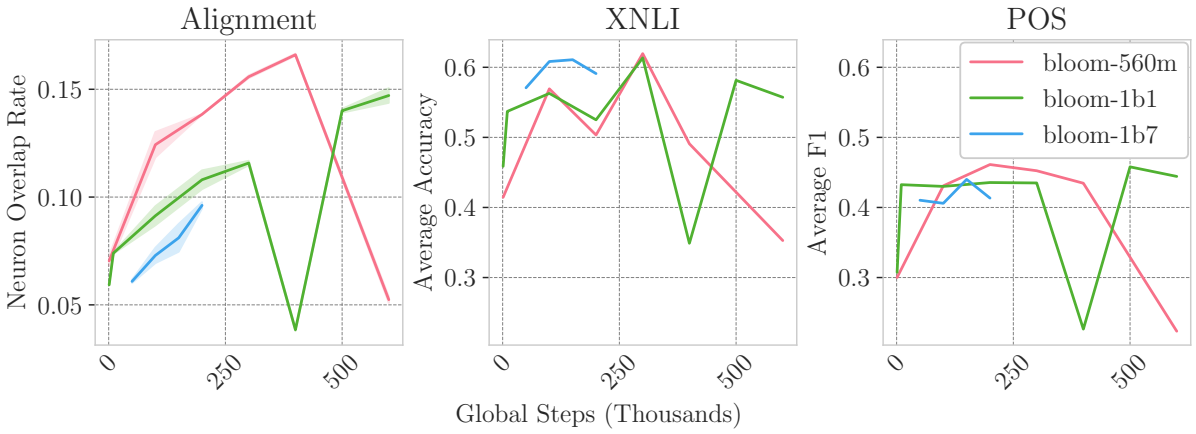


Figure 3: **Left**: The trend of neuron overlap rates (averaged between layers 13 and 17) throughout training. Line colours indicate different model scales. **Centre** and **Right**: Average zero-shot cross-lingual transfer performance across target languages on XNLI and POS tagging.

only requires around 10% of the original model parameters, and each model could be fine-tuned on a single 80GB NVIDIA A100 GPU.[4]

4 Results

4.1 The Dynamics of Alignment

The level of cross-lingual alignment during pre-training of BLOOM models is shown in Figure 1. The plots exhibit similar trends across the linguistic properties we probed, within the same model scale; however, they differ significantly across scales. This stands in contrast with our initial assumption of a gradual emergence of language-agnostic representations, which would imply a monotonic increase of neuron overlap. First, we find that the smallest model (BLOOM_{560m}) shows the highest overlap during most of the pre-training steps. Moreover, we notice a dramatic drop of overlap rates in two model scales, which occurs at around 600k global steps for BLOOM_{560m} and a bit earlier at 400k steps for BLOOM_{1b1}. This drop happens at the end of the pre-training of scale 560m, while BLOOM_{1b1} recovers a high rate of neuron overlap in the latter stage of pre-training.

While this phenomenon may be an artefact due to the variance of overlap rates or an error in checkpointing, we remark that similar drops were also observed in encoder-only multilingual LMs. In fact, Blevins et al. (2022) also detects a performance degradation point among a series of XLM-R checkpoints when evaluated on dependency relation prediction. This affects both in-language performance and cross-lingual transfer. On the other hand, no similar phenomenon occurs when probing monolingual models: Liu et al. (2021) report that these LMs usually display a steady acquisition of linguistic properties along the pre-training trajectories, retaining high performance maintains after a steep increase at the beginning. In contrast, we find that linguistic features are obtained gradually but inconsistently in multilingual models throughout the pre-training process, as shown in Figure 1.

As BLOOM_{560m}, BLOOM_{1b1}, and BLOOM_{1b7} share the training corpus, hyperparameter setting and architecture, they only differ in model scales; however, only the largest scale, BLOOM_{1b7}, shows a monotonic growth in neuron overlap. Thus, the emergence of cross-lingual alignment might follow a scaling law (Kaplan et al., 2020) only after a certain threshold in model size. This hypothesis is supported by further results discussed in Section 4.3 and could be verified on larger scales, such as BLOOM_{3b} and BLOOM_{7b1}.

4.2 Cross-lingual Transfer Correlation

We also conduct a correlation analysis between the cross-lingual alignment and the zero-shot transfer performance, illustrated in Figure 3. Overall, the zero-shot cross-lingual transfer ability of BLOOM shows a strong correspondence with the neuron overlap throughout pre-training, within each model size. This observation holds true also when considering each target language individually, rather than the cross-lingual average, as shown in Figure 4a and Figure 4b.

[4]To make the results comparable across scales, we apply the same training setting (including QLoRA) to the three sizes of BLOOM.

		XNLI		POS	
		Average	Pairwise	Average	Pairwise
Pearson (r)	BLOOM_{560m}	$0.808_{0.052}$	$\mathbf{0.568}_{8.774e\text{-}05}$	$\mathbf{0.940}_{0.005}$	$\mathbf{0.612}_{3.277e\text{-}09}$
	BLOOM_{1b1}	$\mathbf{0.804}_{0.016}$	$\mathbf{0.723}_{3.081e\text{-}10}$	$\mathbf{0.831}_{0.011}$	$\mathbf{0.638}_{1.204e\text{-}12}$
	BLOOM_{1b7}	$0.395_{0.605}$	$\mathbf{0.572}_{0.001}$	$0.258_{0.742}$	$\mathbf{0.534}_{2.691e\text{-}05}$

Table 1: Correlation analysis on average (shown by Fig. 3) and pairwise (shown by Fig. 4a and 4b) overlap rate and zero-shot transfer performance by Pearson coefficient, where the p-values are displayed as subscripts. Colours of p-values indicate statistical significance ($p < 0.05$), high statistical significance ($p < 0.001$) and no statistical significance ($p \geq 0.05$). Coefficients larger than 0.5 with significance under the null hypothesis are bold.

We measure the strength of the correlation in terms of Pearson's coefficients r and its statistical significance against the null hypothesis as p-values. As shown in Table 1, we compute the correlations on the average (Figure 3) and pairwise (Figure 4a and Figure 4b) neuron overlap for each model size against both XNLI and POS tagging performance. The correlations are noticeably higher overall for pairwise measurements (as opposed to average metrics) and for the two smaller models (560m and 1b1). Nonetheless, the fact that pairwise correlations for all model scales are both strong and significant lends credibility to our claim that neuron overlap is tightly connected with zero-shot cross-lingual transfer abilities. Thus, we verify that multilingual LMs transfer between languages more easily if more shared neurons are aligned while pretraining.

4.3 Why does the drop point occur?

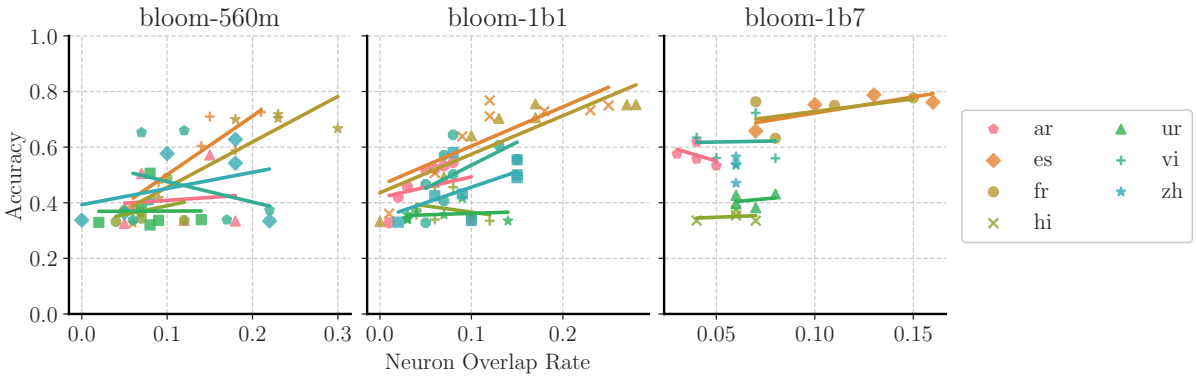
As already mentioned in §4.1, we observe that an unexpected drop of cross-lingual alignment may appear during the training process, suggesting that multilingual LMs can pass through highly sub-optimal regions in the loss landscape. As depicted by the drop points in Figure 1 and Figure 3, intermediate models saved on global step *600k* of BLOOM_{560m} and *400k* of BLOOM_{1b1} have common traits: **1)** there is nearly no neuron overlap detected in these models; **2)** These models display weak zero-shot transfer abilities since their performance on target languages are mostly random, as shown in Figure 5; **3)** These models' performance becomes worse also in-language, showing a performance degradation in the source language English.

Several works studying linguistic acquisition in LMs across time find that the risk of models falling into bad minima depends on their scale. Xia et al. (2023) conduct a study on the training dynamics across scales in monolingual LMs. They find that while all models *decrease* their perplexity for hallucinated texts at the start of training, only large-scale models eventually escape this sub-optimal distribution. Conneau et al. (2020a) introduce the *curse of multilinguality*, which refers to the phenomenon that given a fixed number of parameters, continued increase in the number of languages leads to a performance degradation in terms of both monolingual and cross-lingual skills. Consequently, they argue that this bottleneck could be solved by increasing model scales.

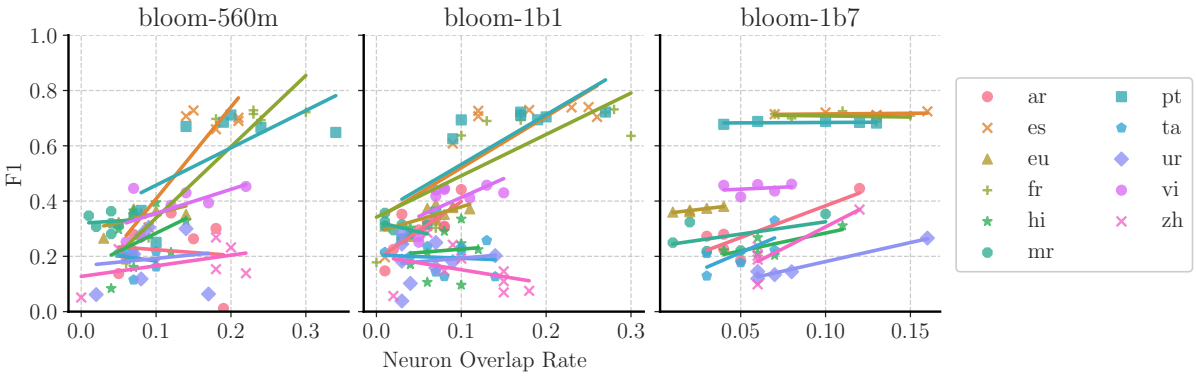
Our experimental results on the dynamics of cross-lingual alignment are in agreement with the aforementioned works. We suggest that cross-lingual alignment follows the same trajectory of learning across scales, similar to what is observed in Chen et al. (2024). However, smaller scales appear more likely to pass through or fall into sub-optimal parameter configurations, which lead to a simultaneous degradation in both in-language and cross-lingual abilities. Our work thus offers a more nuanced perspective on the multilingual abilities of LMs based on learning dynamics, which enriches the received wisdom based on converged models.

5 Related work

Probing Linguistic Features in Multilingual LMs. Probing is a prevalent approach for model interpretability, which is used to examine the information encapsulated in the hidden representation of LMs (Taktasheva et al., 2021; Papadimitriou et al., 2021), including multilingual LMs such as m-BERT and XLM (Lample and Conneau, 2019). Previous work demonstrated that embedding spaces in different languages tend to be isomorphic, and can be better aligned *post-hoc* with the aid of parallel examples or anchor points, which improves zero-shot cross-lingual performance (Cao et al., 2020; Schuster et al., 2019; Conneau et al., 2020b).



(a) Pairwise cross-lingual correlation on XNLI.



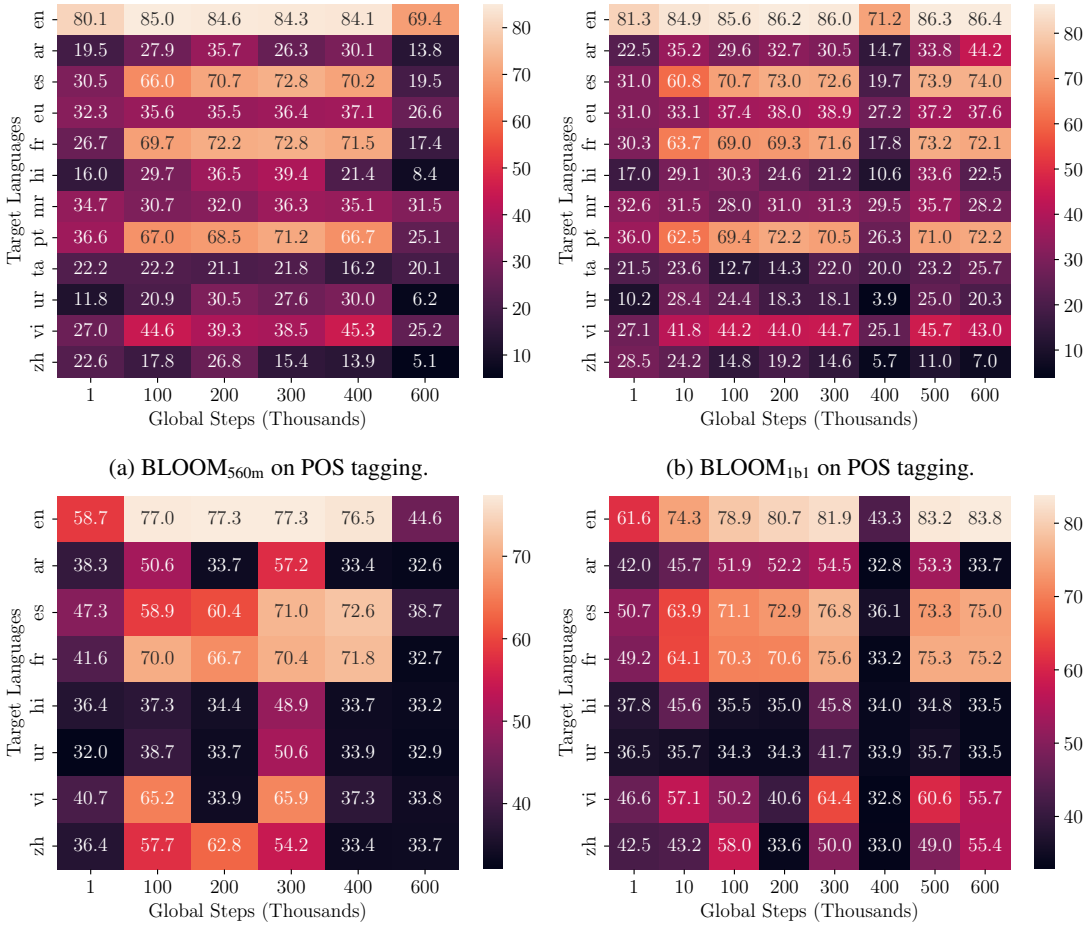
(b) Pairwise cross-lingual correlation on POS.

Figure 4: Neuron overlap rate, which measures the extent of cross-lingual alignment, plotted against the zero-shot cross-lingual transfer performance on (a) XNLI and (b) POS tagging for all checkpoints. The colours indicate the target languages paired with English. The Pearson's correlation coefficients are shown in Table 1.

Structural Overlap and Generalisations. Overlap in neurons (dimensions of hidden representations) or subnetworks of parameters are considered to support generalisation abilities. A direction of research attempts to identify language-specific neural subnetworks, finding that they are topologically similar (Foroutan et al., 2022) and that their overlap might depend on typological distance (Ansell et al., 2022, 2023). In addition, Muller et al. (2021) detected a high correlation between the similarity of representations and the zero-shot cross-lingual transfer performance in the converged mBERT. All of these works imply a strong correlation between cross-lingual alignment and zero-shot transfer ability, which is further confirmed by our study on the training trajectory of BLOOM across scales.

Recently, Bhaskar et al. (2024) finds that all competing subnetworks within LLMs, which have similar in-domain performance but different out-of-domain generalisation, share a so-called 'heuristc core', while Templeton et al. (2024) demonstrated that sparse auto-encoders can identify various features—from a specific landmark to code errors—in a production-grade LLM. Our work, in conjunction with theirs, provides a reliable framework for explaining model generalisation through the lens of shared neurons.

Knowledge Acquisition during Pre-training. Concurrently, there is a rising interest in understanding the training dynamics of LLMs. Works that mainly examine monolingual English models report a steady trend in the acquisition of linguistic knowledge. Both Xia et al. (2023) and Choshen et al. (2022) argue that language acquisition undergoes the same order of phase transitions consistently across model scales, training objectives and random seeds. Chen et al. (2024) find that the emergence of syntactic structure in the attention scores of Transformer-based LMs is essential for grammar acquisition in LMs, but does not account for semantic knowledge acquisition. For multilingual training, Choenni et al. (2023) examine how data size and language variance affect the performance during fine-tuning. The experiments presented by Blevins et al. (2022) are the most reminiscent of



(a) BLOOM_{560m} on POS tagging.

(b) BLOOM_{1b1} on POS tagging.

(c) BLOOM_{560m} on XNLI.

(d) BLOOM_{1b1} on XNLI.

Figure 5: The zero-shot cross-lingual performance of BLOOM checkpoints of size 560m (left) and 1b1 (right) on POS tagging (top) and XNLI (bottom).

our work. They focus on the inconsistency between the emergence of in-language and cross-language abilities for encoder LMs, whereas we study the dynamics of neuron overlaps and the corresponding impact on downstream performance in autoregressive LMs.

6 Conclusions

In this paper, we probe a collection of checkpoint models of BLOOM to study the dynamics of multilingual pretraining. By experimenting with three model sizes, we observe that the subset of neurons encoding linguistic features tends to increase their overlap across languages throughout pretraining. Nevertheless, we also detect severe drops that occur at different points in the training process, especially at smaller model scales, instead of a steady increase in the extent of alignment.

Moreover, we corroborate the hypothesis that the shared neurons are tightly connected with the zero-shot cross-lingual transfer ability of multilingual LLMs: the same sub-networks are activated at inference time and updated during fine-tuning, which contributes to the cross-lingual generalisation ability of LMs. This assumption is further confirmed across model scales by observing a high correlation between neuron overlap and downstream task performance in syntactic and semantic tasks. Hence, our work contributes to understanding how multilingual LMs implicitly align information across languages even in the absence of parallel data.

Limitations

Our work focuses on the checkpoints of BLOOM with large intervals in global steps. Thus, our findings on the trend of alignment might be not applicable if zooming in on a particular window of training with finer-grained checkpoint models. Moreover, we consider only autoregressive models with the same objective and training dataset: varying these

properties may result in different patterns.

Although many of our findings on the dynamics of cross-lingual alignment align with previous research on encoder Transformers, some aspects of the experimental design (e.g., selected layers and morphosyntactic categories) are not directly transferable to other architectures or training corpora. Moreover, we focus on the alignment of languages seen during pretraining, whereas the generalisation to unseen languages is left for future research.

Acknowledgements

This work used resources provided by the Edinburgh Compute and Data Facility (ECDF).[5]

References

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2023. Distilling efficient language-specific models for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8147–8165, Toronto, Canada. 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*, pages 4623–4637, Online. Association for Computational Linguistics.

Adithya Bhaskar, Dan Friedman, and Danqi Chen. 2024. The heuristic core: Understanding subnetwork generalization in pretrained language models.

BigScience Workshop, 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

[5] http://www.ecdf.ed.ac.uk/

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, Daniel McDuff, 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 Nezhurina, 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. 2023. BLOOM: A 176b-parameter open-access multilingual language model.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during LM fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13244–13257, Singapore. Association for Computational Linguistics.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language*

Processing, pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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, Gerard 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 Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS Corpus: A 1.6TB composite multilingual dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual BERT?

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.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Karolina Stańczak, Lucas Torroba Hennigen, Adina Williams, Ryan Cotterell, and Isabelle Augenstein. 2023. A latent-variable model for intrinsic probing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13591–13599.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 191–210, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738, Toronto, Canada. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Languages List

A.1 Target Languages List

Based on our hypothesis introduced in Section 1, we select the set of target languages from the intersection of treebanks in UD v2.1 and the BigScience ROOTS Corpus (Laurençon et al., 2022) used for pre-training BLOOM, so that we can perform implicit alignment detection following the procedure described in §3.1. In addition to this criterion, we further select target languages based on the data availability for experiments on downstream tasks in Section 3.2. A full list is given below.

UD v2.1	ar	eu	ca	zh	en	fr	hi	mr	pt	es	ta	ur	vi
§3.2 XNLI	✓			✓	✓	✓	✓			✓		✓	✓
§3.2 POS	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

A.2 Language and Morphosyntactic categories

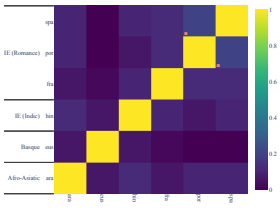
The following table lists the corresponding morphosyntactic categories for the languages we probed.

Language	ISO 639-1 code	ISO 639-3 code	Aspect	Case	Definiteness	Finiteness	Gender	Mood	Number	Person	POS	Tense	Voice
Arabic	ar	ara	✓	✓	✓		✓	✓	✓		✓		✓
Basque	eu	eus	✓	✓	✓				✓		✓		
Catalan	ca	cat					✓	✓	✓	✓	✓	✓	
Chinese	zh	zho									✓		
English	en	eng							✓	✓	✓	✓	
French	fr	fra	✓				✓		✓	✓	✓	✓	
Hindi	hi	hin	✓	✓		✓	✓	✓	✓	✓	✓		✓
Marathi	mr	mar					✓		✓	✓	✓		
Portuguese	pt	por	✓				✓	✓	✓	✓	✓	✓	
Spanish	es	spa	✓				✓	✓	✓	✓	✓	✓	
Tamil	ta	tam		✓		✓	✓		✓		✓	✓	
Urdu	ur	urd		✓		✓			✓		✓		
Vietnamese	vi	vie									✓		

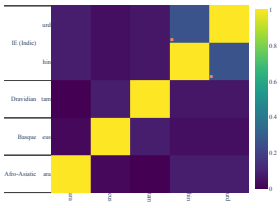
B Results

B.1 Pairwise Overlap Comparison

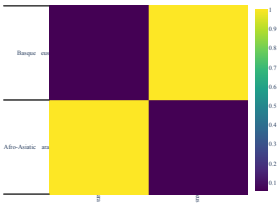
In this section, we exhibit an exhaustive collection of heatmaps of layer 17 in the converged $\text{BLOOM}_{560\text{m}}$ for all the possible morphosyntactic categories listed in Appendix A.2. The orange dot indicates an overlap that is statistically significant under the null hypothesis.



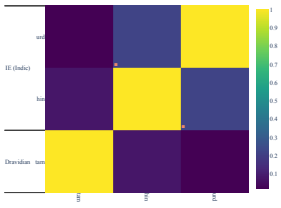
(a) Aspect



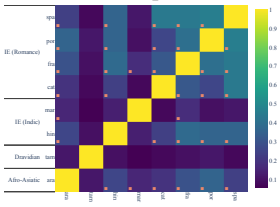
(b) Case



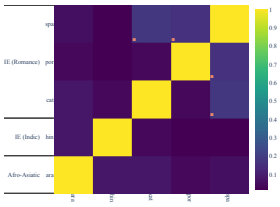
(c) Definiteness



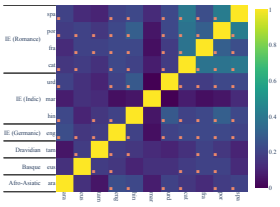
(d) Finiteness



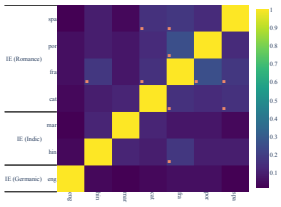
(e) Gender



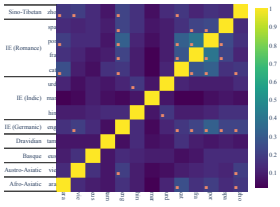
(f) Mood



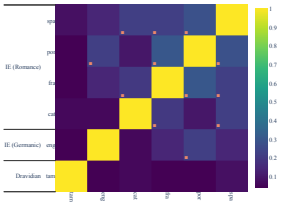
(g) Number



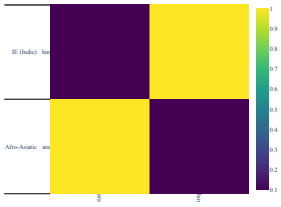
(h) Person



(i) POS



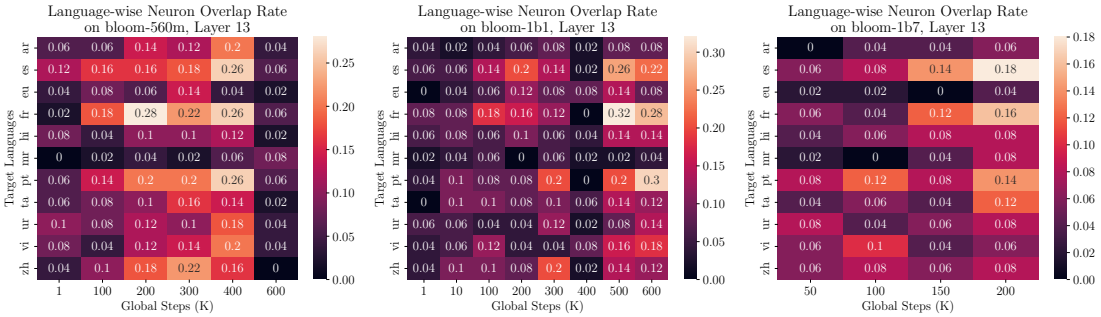
(j) Tense



(k) Voice

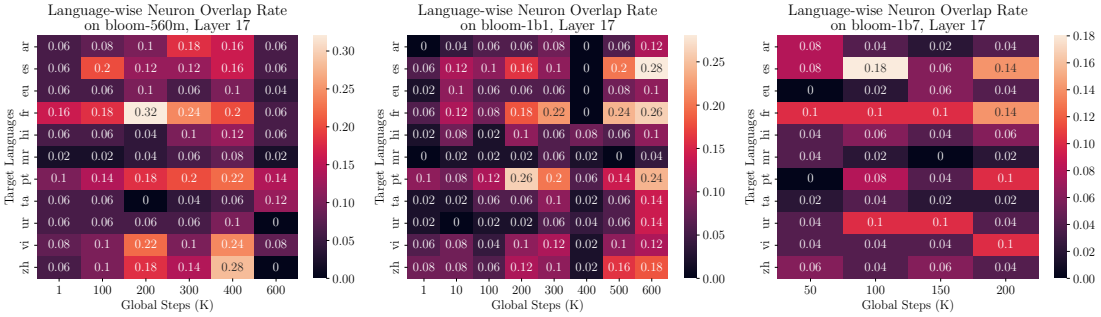
B.2 Pairwise overlap rates throughout training

The neuron overlap rate between target languages and English in Layer 13:



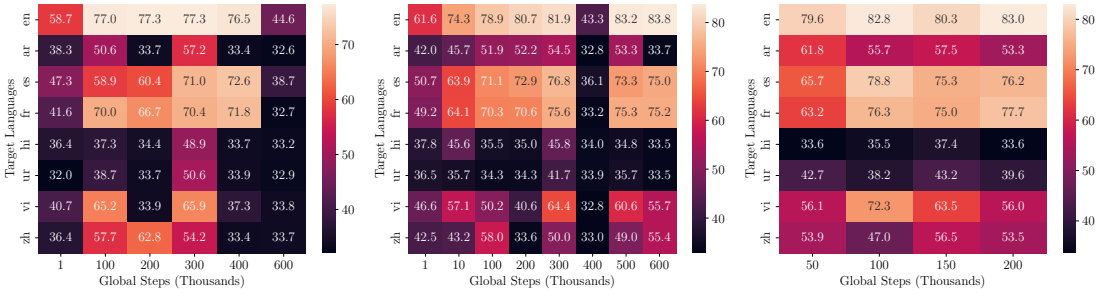
(a) bloom-560m (b) bloom-1b1 (c) bloom-1b7

The neuron overlap rate between target languages and English in Layer 17:

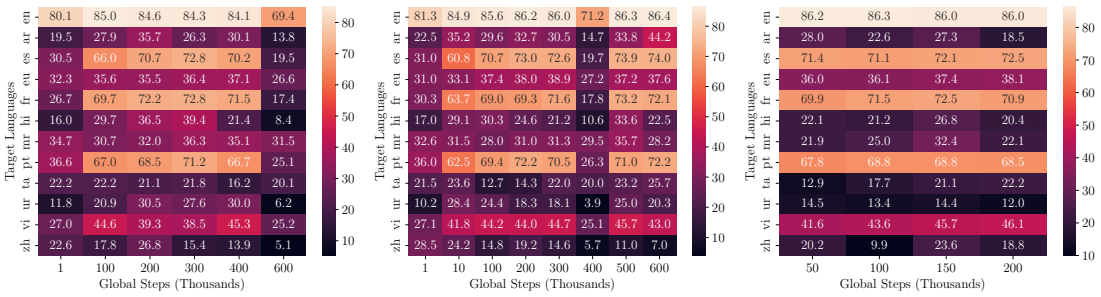


(a) bloom-560m (b) bloom-1b1 (c) bloom-1b7

B.3 Zero-shot Cross-lingual Performance on Downstream Tasks



(a) bloom-560m, XNLI (b) bloom-1b1, XNLI (c) bloom-1b7, XNLI



(a) bloom-560m, POS tagging (b) bloom-1b1, POS tagging (c) bloom-1b7, POS tagging