Cross-Lingual Generalization and Compression: From Language-Specific to Shared Neurons

Frederick Riemenschneider and **Anette Frank**
Department of Computational Linguistics
Heidelberg University, Germany
{riemenschneider|frank}@cl.uni-heidelberg.de

Abstract

Multilingual language models (MLLMs) have demonstrated remarkable abilities to transfer knowledge across languages, despite being trained without explicit cross-lingual supervision. We analyze the parameter spaces of three MLLMs to study how their representations evolve during pre-training, observing patterns consistent with compression: models initially form language-specific representations, which gradually converge into cross-lingual abstractions as training progresses. Through probing experiments, we observe a clear transition from uniform language identification capabilities across layers to more specialized layer functions. For deeper analysis, we focus on neurons that encode distinct semantic concepts. By tracing their development during pre-training, we show how they gradually align across languages. Notably, we identify specific neurons that emerge as increasingly reliable predictors for the same concepts across languages. This alignment manifests concretely in generation: once an MLLM exhibits cross-lingual generalization according to our measures, we can select concept-specific neurons identified from, e.g., Spanish text and manipulate them to guide token predictions. Remarkably, rather than generating Spanish text, the model produces semantically coherent English text. This demonstrates that cross-lingually aligned neurons encode generalized semantic representations, independent of the original language encoding.

1 Introduction

How do multilingual language models achieve cross-lingual generalization? This question has puzzled researchers for years–particularly since standard pre-training objectives do not explicitly encourage cross-lingual alignment.

Existing research has explored various potential explanations, ranging from linguistic similarity due to genetic or geographic relatedness (Lin et al., 2019; Lauscher et al., 2020) and word order properties (Dufter and Schütze, 2020; Deshpande et al., 2022) to architectural features like shared subwords (Pires et al., 2019; Wu and Dredze, 2019, *inter alia*). However, researchers found conflicting evidence that challenges these explanations, especially regarding the role of shared subwords (Artetxe et al., 2020; K et al., 2020, *inter alia*).

In this paper, we aim to derive explanations at a deeper level, by exploring the connection between cross-lingual generalization and signals of compression. We build on the theory that the limited capacity of language models forces them to discover efficient, shared representations encoded in specific neurons across languages, rather than maintaining separate, language-dependent encodings.

To gain deeper insights into the development of cross-lingual representations, we focus on the pre-training process itself rather than just analyzing the final model state. For our analysis, we use models from the BLOOM family (BigScience Workshop, 2022) at different scales (BLOOM-560M and BLOOM-7B1), which are state-of-the-art multilingual decoder models that provide open access to training checkpoints, though their number is limited. Additionally, we introduce our own decoder-only model where we collect checkpoints at much finer intervals and maintain precise control over the training data and languages, allowing us to build a detailed picture of how cross-lingual representations are shaped during training.

While much prior work has focused on zero-shot cross-lingual transfer performance, recent research has shown such transfer to be unreliable (Rajaee and Monz, 2024). We therefore take a more mechanistic approach and analyze the models themselves rather than their zero-shot transfer capabilities. We start our analysis by probing model-internal representations for language identity prediction, to build initial intuitions, which reveals clear shifts in performance across pre-training checkpoints. We

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13470–13491
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

then examine how semantic concepts (like house or earthquake) are represented in *individual neurons* across different languages, following the neuron analysis approach of Suau et al. (2022).

Our analysis reveals increasing cross-lingual alignment during pre-training, with specific neurons emerging as shared concept experts across languages. We complement these observations with an information-theoretic perspective on compression and demonstrate their practical implications through controlled text generation experiments, showing how MLLMs evolve from language-specific to generalized neural representations during the pre-training process. To summarize, our contributions are:

(i.) We provide empirical evidence for the compression hypothesis in MLLMs by tracking how representations evolve from language-specific to cross-lingual throughout training, using mechanistic interpretability methods and special probing tasks.

(ii.) To our knowledge, we are the first to analyze the development of cross-lingual semantic generalization during pre-training, by identifying specific neurons that encode the same concepts across different languages. Our analysis reveals how semantic information concentrates in middle layers and evolves into generalized concept representations shared across languages at later training stages.

(iii.) We demonstrate how our findings have a visible effect on text generation through controlled neuron manipulation experiments, illustrating that the model's internal representations encode shared conceptual knowledge beyond specific language boundaries.

We release our MLLM with comprehensive training checkpoints, along with code and data.[1]

2 Related Work

Studying Cross-Lingual Generalization. Since the early development of MLLMs, researchers have investigated their remarkable effectiveness, primarily through the lens of zero-shot cross-lingual transfer performance. Explanatory factors for the identified cross-lingual generalization capabilities can roughly be divided into *linguistic aspects*, such as genetic and geographic relatedness and word order (Lin et al., 2019; Lauscher et al., 2020; Dufter and Schütze, 2020; Deshpande et al., 2022, *inter alia*), and *architectural considerations* like model depth and number of parameters (Dufter and Schütze, 2020; K et al., 2020, *inter alia*). The role of *lexical overlap between languages* has been a particular point of debate in the literature (Pires et al., 2019; Wu and Dredze, 2019; Artetxe et al., 2020; Dufter and Schütze, 2020; K et al., 2020, *inter alia*). For a comprehensive overview of these developments, we refer to Philippy et al. (2023).

Compression in Language Models. The information bottleneck method, introduced by Tishby et al. (1999), provides a theoretical framework for analyzing information flow in neural networks. Tishby and Zaslavsky (2015) apply this framework to deep learning, showing that neural networks must learn to efficiently represent task-relevant information while "forgetting" irrelevant input details. Building on these insights, Voita et al. (2019) analyze how representations evolve bottom-up in Transformers. Shwartz-Ziv and Tishby (2017) identify two distinct phases in neural network training: an *initial fitting phase* followed by a *compression phase*, the latter being causally linked to the network's generalization capabilities.

In the multilingual context, the compression hypothesis has been acknowledged, but to date remains underexplored: Chi et al. (2021) explicitly reference the information bottleneck method while deferring its investigation, and Dufter and Schütze (2020) observe that overparameterization may actually hinder multilingual performance. In our work, we systematically investigate how compression manifests in multilingual models during pre-training, hypothesizing that restricted model capacity forces the development of shared cross-lingual representations, rather than maintaining separate language-specific ones.

Mechanistic Interpretability. Mechanistic interpretability seeks to reverse engineer neural networks to understand their internal functioning. Even in work not focused on MLLMs, multilingual phenomena have emerged as peripheral findings: Gurnee et al. (2023) identify neurons that respond to French texts through sparse probing, while Bricken et al. (2023) discover Arabic script and Hebrew features via sparse autoencoders.

Current work has begun to explicitly address multilinguality: Wendler et al. (2024) show that LLAMA 2 models (Touvron et al., 2023), consistent

[1] https://github.com/Heidelberg-NLP/cross-lingual-generalization.

with their English-dominated training data, process other languages using English as internal pivot. Recently, research has developed specialized methods to identify language-specific neurons: Tang et al. (2024) propose LAPE (Language Activation Probability Entropy), while Kojima et al. (2024) build on Suau et al.'s (2022) methodology, which we introduce below. However, these investigations focus solely on identifying neurons responsible for language-*specific* processing, rather than examining, as we do, whether semantic concepts *share* neural representations across languages.

Closest to our work is Blevins et al. (2022), who analyze XLM-R (Conneau et al., 2020) checkpoints for its performance in linguistic tasks such as PoS tagging or dependency parsing across different languages. Yet, unlike their work, we examine decoder-only models and investigate the pretraining process at a more fundamental level.

3 Conceptualizing Cross-Lingual Generalization

Most prior work studies MLLMs through zero-shot cross-lingual transfer. In this setting, "a model that is fine-tuned on one language can be applied to others without any further training" (Tunstall et al., 2022). This approach has become the standard for evaluating multilingual models (Hu et al., 2020).

Often, zero-shot cross-lingual transfer (often abbreviated to just "*cross-lingual transfer*") is treated as synonymous with *cross-lingual generalization*. This conflation is problematic for two reasons. First, despite its name suggesting otherwise, in zero-shot cross-lingual transfer, models *do* undergo fine-tuning, potentially obscuring more subtle phenomena (Papadimitriou et al., 2023). Second, these evaluations are vulnerable to dataset artifacts like word overlap and answer position bias (Rajaee and Monz, 2024), and may instead reflect surface-level patterns, rather than linguistic generalization.

We therefore argue for a clear distinction between *zero-shot cross-lingual transfer* as a specific evaluation method and *cross-lingual generalization* as the fundamental ability of models to form cross-lingual abstractions. Our work investigates the latter by analyzing internal representations directly, without any fine-tuning. We hypothesize that *cross-lingual generalization emerges through compression* during pre-training. We assume that once a model's capacity constraints prevent pure memorization, it develops more space-efficient representations by abstracting away language-specific features from the encoded content. Our experiments support this hypothesis: we observe the emergence of *cross-lingual concept neurons* that respond to the same concepts across different languages.

4 Model Details

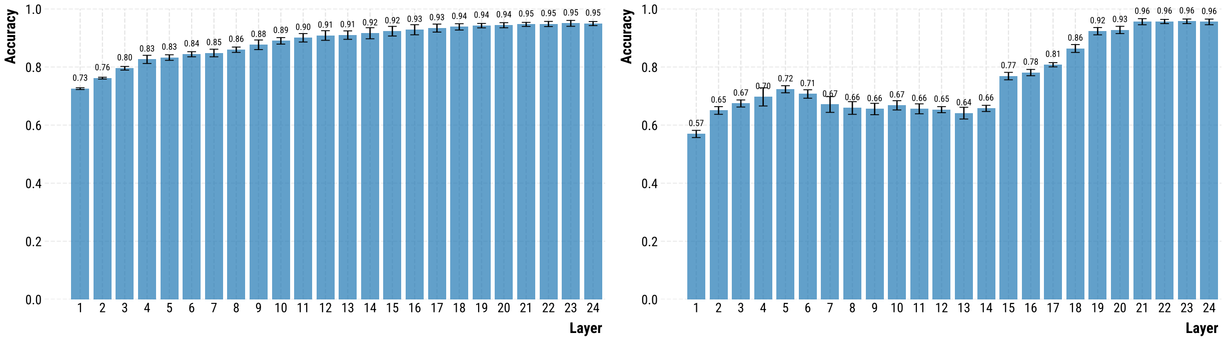
To study cross-lingual generalization during pre-training, we require access to model checkpoints throughout the pre-training process. We therefore focus our analysis on the BLOOM family (BigScience Workshop, 2022), the only recent collection of MLLMs to offer publicly available training checkpoints. Due to computational resource constraints, we conduct our most detailed analysis on BLOOM-560M. However, we confirm our key findings on BLOOM-7B1, demonstrating that our results generalize to larger models.[2]

In addition, to allow for a more fine-grained analysis of training dynamics than BLOOM's checkpoint frequency allows, we pre-trained our own model based on the XGLM architecture (Lin et al., 2022), using a reduced dimension (`d_model` = 512 instead of 1024), resulting in approx. 257M parameters. We trained on 16 languages spanning diverse language families and scripts (Germanic, Italic, Bantoid, and Slavic), collecting checkpoints at powers of two and regular 5000-step intervals.

5 Probing Language Identity Across Layers and Checkpoints

Motivation. To investigate the relationship between cross-lingual generalization and the model's representations, we examine to what extent language-specific information is encoded across layers and training stages, focusing on how the MLLM's internal organization of languages develops during training. In a first step, we probe each model layer's ability to identify which language is being processed, examining how language-specific information is distributed across the model's layers. This initial probing experiment serves as a foundation for understanding how language representations develop both through the model's layers and throughout its training process.

[2] Two checkpoints each from BLOOM-560M and BLOOM-7B1 were excluded from our analysis, as they appear to be corrupted (cf. https://huggingface.co/bigscience/bloom-560m-intermediate/discussions/2 and https://huggingface.co/bigscience/bloom-7b1-intermediate/discussions).



(a) Early training stage (step 1000).

(b) Late training stage (step 400 000).

Figure 1: Language identity probing classification accuracy across layers of the BLOOM-560M model at different training stages. Higher accuracy indicates that language-specific information is more easily extractable from the hidden states at that layer. Error bars show standard deviation across three random seeds.

Experiment Setup. For a given language l, we sample $\{s_0^l, s_1^l, ..., s_n^l\}$ sentences from the OSCAR corpus (Ortiz Suárez et al., 2019). Each sentence s_i^l is tokenized into a sequence of tokens $[t_{i,0}^l, t_{i,1}^l, ..., t_{i,T_i-1}^l]$. For each tokenized sentence, the model $\mathcal{M}$ produces hidden representations: $h_{i,0}^l, h_{i,1}^l, ..., h_{i,T_i-1}^l = \mathcal{M}(t_{i,0}^l, t_{i,1}^l, ..., t_{i,T_i-1}^l)$.

From these sequences of hidden states, we randomly sample one token position per sentence and extract the hidden representation at that position. For instance, for sentence s_i^l, we might select position p_i^l to obtain $h_{i,p_i^l}^l$. We then train a logistic regression classifier on these sampled hidden states, aiming to predict which language l the hidden state originated from. By analyzing classification performance across layers, we investigate how the representation of languages evolves throughout the MLLM's architecture, and how languages are organized. For implementation details see Appendix B.

Results. We present analysis results for BLOOM-560M at pre-training steps 1000 and 400 000 in Figure 1. At step 1000, the model already demonstrates strong language identification capabilities, with a slight performance increase after the first layer followed by small, monotonic improvements across subsequent layers. At step 400 000, by contrast, we observe markedly different behavior: performance in earlier layers is substantially weaker, starting at 57% accuracy in the first layer, increasing until layer 5, then declining until layer 14. From layer 15 onwards, performance recovers, eventually matching the levels observed in the earlier checkpoint.

The precise layer-wise accuracy trajectory appears to be architecture-dependent, with decoder-only models showing this distinctive pattern (see results for BLOOM-7B1 and our toy model, as well as comparisons with encoder-only models XLM-R (Conneau et al., 2020) and MBERT (Devlin et al., 2019) in Appendix B). However, we observe a fundamental organization that is shared across different model families: language-specific information diminishes in the middle layers, while the final layers maintain strong identification capabilities.

Complementing our detailed analysis of individual checkpoints in Figure 1, Figure 2 tracks three key statistics throughout training: the first layer accuracy, the mean probing accuracy averaged across all layers, and the corresponding standard deviation between layer-wise accuracies. During early training (steps 1000 to 10 000), we observe uniformly high language identification performance across layers, reflected in high mean accuracy and low between-layer variance. Beyond step 100 000, layer-averaged accuracy decreases (especially in the first layer), while standard deviation increases, indicating greater differentiation between layers.

These findings reveal a fundamental shift in how language information is processed throughout pre-training: the model initially develops strong language identification capabilities, but subsequently this ability diminishes. We hypothesize that different layers develop distinct functional roles during training: while final layers maintain high language identification accuracy necessary for next-token prediction, middle layers develop representations that tend to be more language-agnostic. These observations can be argued to provide initial evidence for a compression effect that manifests during pre-training, characterized by a shift from language-specific to more generalized representations.

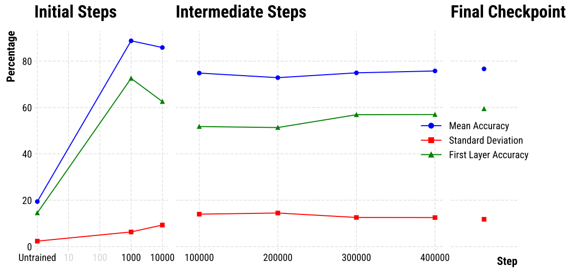


Figure 2: Language identification probing accuracy throughout training of BLOOM-560M. For each checkpoint we show: (1) mean accuracy across layers, (2) standard deviation across layers (indicating how much accuracy varies between layers), and (3) first layer accuracy, which exhibits the most significant changes during training. Results averaged over three random seeds.

In what follows, we build on this hypothesis and investigate whether cross-lingual generalization can be explained as a result of *compression* that occurs in a multilingual model's representation space after an initial phase of memorization.

6 Tracing Concepts in Neurons

Research Question. We now examine how individual concepts are represented in MLLMs by tracing their evolution during pre-training, examining their relationships across languages. We hypothesize that MLLMs first develop language-specific representations of concepts (e.g., separate encodings for "house", "casa", or "Haus"), which, as training progresses, merge into unified abstractions (e.g., the concept of a `dwelling`). We expect that these abstracted representations can be "projected" into language-specific instantiations during generation, offering a more space-efficient organization than separate representations for each language.

Experiment Setup. To identify concept-specific "expert" neurons (e.g., those specialized in representing `dwelling`), we adopt the methodology of Suau et al. (2022). Each concept c in language l is represented by a dataset $\{\mathbf{x}_i^{c,l}, b_i^{c,l}\}_{i=1}^N$, where the total of $N_{c,l} = N_{c,l}^+ + N_{c,l}^-$ sentences are divided into positive samples that contain the concept ($b_i^{c,l} = 1$) and negative ones that do not ($b_i^{c,l} = 0$). A neuron demonstrates *expertise* for concept c if it selectively activates for positive examples, while remaining inactive for negative ones.

We evaluate how well an MLP neuron m's activation pattern (excluding attention neurons) predicts concept c by analyzing the neuron's outputs $\mathbf{z}_m^{c,l} = \{z_{m,i}^{c,l}\}_{i=1}^N$ in response to sentences $\{\mathbf{x}_i^{c,l}\}$ and use these activations as *concept prediction scores*, that indicate the presence of concept c in a given input for language l.[3] We measure a neuron's predictive power through Average Precision $\text{AP}_m^{c,l} = \text{AP}(\mathbf{z}_m^{c,l}, \mathbf{b}^{c,l})$, which quantifies the area under the precision-recall curve.

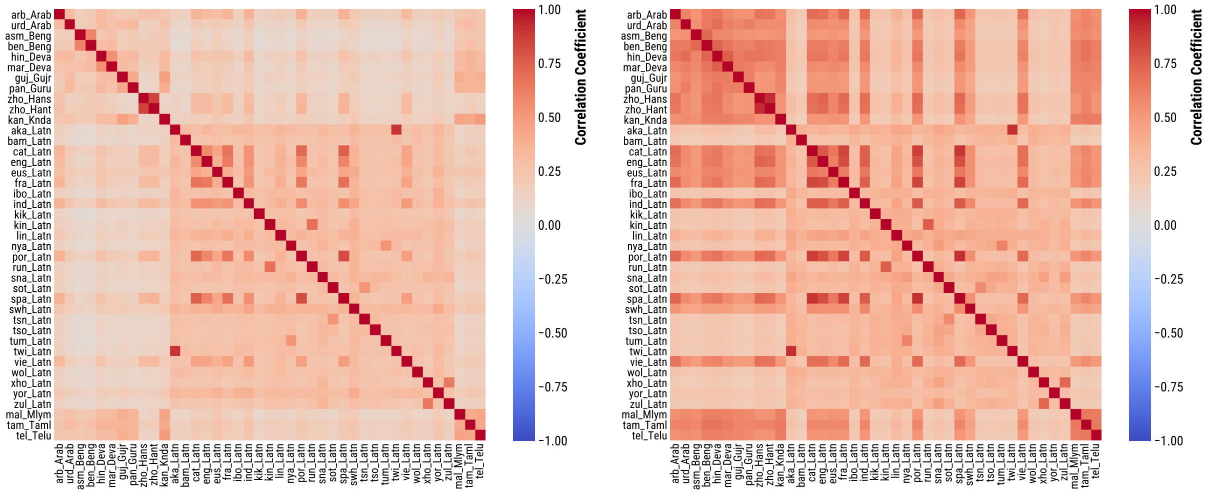
Data. We follow Suau et al. (2022) in constructing our concept dataset from OneSec (Scarlini et al., 2019), which provides Wikipedia sentences annotated with WordNet senses (Miller, 1994). From this corpus, we sample 200 WordNet senses as target concepts, ensuring $100 \leq N_{c,\text{eng}}^+ \leq 1000$ positive samples and $N_{c,\text{eng}}^- = 1000$ negative samples per concept. We translate the resulting English dataset N_{eng} using the NLLB 1.3B model (Costa-jussà et al., 2022) to create parallel versions in the languages we use in our analysis. Details on the dataset construction process are given in Appendix C.

General Neuron Alignment. Using our multilingual corpus derived from OneSec and the concept prediction score introduced above, we compute, for each language l and concept c, an *expert score vector* $\mathbf{e}^{c,l} \in \mathbb{R}^M$, where M is the number of neurons and each element $e_m^{c,l} = \text{AP}_m^{c,l}$ represents the expertise of neuron m for concept c in language l. To investigate whether *the same neurons* specialize in representing the same concepts across languages, we analyze the *cross-lingual alignment* of these expert scores. Specifically, we compute the Pearson correlation coefficient between expert score vectors $\mathbf{e}^{c,l_1}$ and $\mathbf{e}^{c,l_2}$ for each concept across different language pairs (l_1, l_2).

Given the large number of pairwise correlations across concepts and languages, we need a way to summarize these results concisely. We therefore apply Fisher's Z transformation to the correlation coefficients, compute their average, and transform the result back. We emphasize that this averaged score cannot be interpreted as a statistical correlation, but it still serves as a meaningful indicator of the degree of neuron alignment between languages.

The resulting matrices for BLOOM-560M at training steps 1000 and 400 000 are shown in Figure 3, with values averaged across all concepts. Figure 4 provides a view of this alignment throughout the training process, showing the averaged scores across both concepts and language pairs. For addi-

[3] A fixed-size sentence representation is obtained via max-pooling.



(a) Early training stage (step 1000).

(b) Late training stage (step 400 000).

Figure 3: Expert neuron alignment across languages in BLOOM-560M at different training stages, measured by Pearson correlation coefficients averaged across concepts using Fisher's Z transformation.

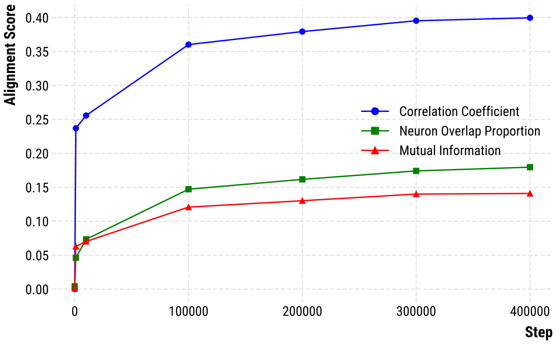


Figure 4: (1) Correlation Coefficient, (2) Neuron Overlap Proportion of top 500 neurons, and (3) MI throughout training, averaged across concepts and languages for BLOOM-560M.

tional matrices and results for BLOOM-7B1 and our toy model see Appendix D.

Early in training, the alignment between languages is relatively weak, but it strengthens substantially by step 400 000. The correlation matrix reveals dependencies that are partially attributable to script families. We observe small but distinct clusters of related languages sharing the same script (e.g., Assamese and Bengali, Hindi and Marathi), and a broader positive alignment across languages using the Latin alphabet. Most notably, there appears to be a strong distinction between Latin-script and non-Latin-script languages, though this pattern is not absolute. The Dravidian languages (Kannada, Malayalam, Tamil, Telugu) represent a case of a language family exhibiting high similarities despite using distinct scripts, reinforcing previous findings that subword overlap alone cannot explain cross-lingual generalization. A deeper analysis of these relationships remains for future investigation.

Information-Theoretic Perspective. Beyond examining correlations across neurons, we adopt an information-theoretic approach by analyzing the Mutual Information (MI) between neural representations across languages. MI quantifies how much knowledge of a concept's representation in one language informs its representation in another language. Specifically, MI measures *compression efficiency* by indicating to what degree a concept's representation in one language is redundant, and thus predictable, given its representation in another language. We compute MI for continuous data using entropy estimation based on k-nearest neighbors distances, following the methods of Kraskov et al. (2004) and Ross (2014), as implemented in scikit-learn (Pedregosa et al., 2011). The evolution of MI (Figure 4) closely mirrors the correlation-derived alignment scores, reinforcing our findings through an information-theoretic lens.

Neuron Overlap. Finally, we analyze the concrete overlap between the most concept-selective neurons across languages. For each concept c and language l, we identify the set $S^{c,l}$ of the top k neurons with the highest expertise scores $\mathbf{e}^{c,l}$. We quantify the cross-lingual overlap between languages l_1 and l_2 using the overlap proportion $O^c_{l_1,l_2} = \frac{|S^{c,l_1} \cap S^{c,l_2}|}{k}$. This directly measures the degree of neuron sharing between languages, suggesting compression, as shared neurons indicate a more compact representation of concepts.

Figure 4 shows the resulting overlap for $k = 500$, averaged across languages and concepts. The evolution of the overlap aligns with both the correlation-derived and mutual information measures. Remarkably, among the more than 200 million MLP parameters analyzed, we find that approx. $\frac{1}{6}$ of the top 500 concept-selective neurons are shared between any pair of languages. This substantial overlap, despite the model's vast capacity, suggests significant cross-lingual representation sharing.

7 Revisiting Layer Distributions

As shown in Section 5, early layers partially lose their language identity information during pretraining. We now return to this observation and examine how it relates to the distribution of concepts across a model's layers. Specifically, we investigate where concept-specific neurons are located, and how their distribution evolves during training.

Layer-Wise Distribution of Expert Neurons. First, we explore where the previously identified top k expert neurons are located across layers. By examining the layer distribution of these neurons, averaged across all languages and concepts, we obtain a language-agnostic view of where concept-specific information is concentrated in the model.

Our analysis in Figure 5 reveals how the concentration of concept information across layers evolves throughout training. In the randomly initialized model, the first layer contains the highest concentration of expert neurons. This is intuitive, as the untrained model can only use surface-level word overlap to "identify" concepts. This first-layer dominance intensifies during early training (step 1000), suggesting that the model initially relies heavily on these lexical cues.

At step 100 000 we observe a fundamental reordering of concept information across layers, which stabilizes and shows only marginal changes in later checkpoints. This new distribution reveals three distinct regions: After initial concentration in the first layers, there is a notable drop reaching its lowest point at layer 10. This same layer marks the beginning of the first of two concentration peaks. The final layer (24) shows a particularly low proportion, likely due to its role in token generation.

Layer-Wise Cross-Lingual Semantic Overlap. Building on these insights, we now analyze the *cross-lingual alignment of concept representations for different layers*. For each layer ℓ, we compute the average pairwise overlap between top k expert neurons by comparing the sets $S_\ell^{c,l}$ across languages, measuring the proportion of shared neurons between languages l_1 and l_2 as $\frac{|S_\ell^{c,l_1} \cap S_\ell^{c,l_2}|}{k}$.

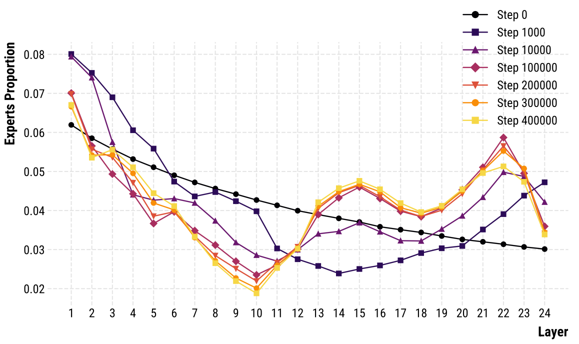


Figure 5: Layer-wise distribution of BLOOM-560M's top 500 expert neurons, averaged across languages and concepts.

The results in Figure 6 complement our previous findings (Figure 5). While the early layers showed high concept specificity, likely due to subword overlap, this does not lead to strong cross-lingual alignment. Instead, substantial cross-lingual overlap develops in the middle layers (10-17), particularly in later checkpoints, suggesting that genuine semantic generalization occurs in this region. This indicates that while subword similarity provides a useful initial bias for the model, actual cross-lingual semantic representations emerge in the middle layers. The increasing alignment in later checkpoints supports the compression hypothesis, suggesting that the model learns to abstract away from language-specific features. The decrease in overlap in the final layers aligns with their specialization for language-specific token generation. By comparing Figure 6 and Figure 5, we can disentangle the effect of subword overlap from true cross-lingual generalization. We confirm the same trends for BLOOM-7B1 in Figures 16 and 17.

8 Steering Text Generation

Until now, we have examined probing performance and neuron behavior to understand neuron alignment in MLLMs. We now investigate whether these findings are reflected in BLOOM-560M's text generation capabilities. To test the cross-lingual semantic properties of concept-specific expert neurons, we adapt the neuron manipulation technique from Suau et al. (2022). For a given

Checkpoint	Concept	Language	Generation
10 000	earthquake	Spanish	Posteriormente se quem se hizo sentir en el segundo momento una intensidad máxima de unos 40 minutos y la presencia de varios volcanes. [...]
		Simplified Chinese	去年,大连市高岭土场镇发现30余处安全隐患。经市安全气象台和地质灾害防御站队员检查,发现大量高空存在安全隐患, [...]
	joy	Spanish	Por todo lo que he leído sobre este nuevo reto, me ha encantado y he querido brotar las historias [...]
		Simplified Chinese	你越长大越幸福,幸福带给你的就是一生的幸福。你越长大越幸福,幸福带给你的就是一生的幸福。 [...]
400 000	earthquake	Spanish	Strong earthquakes occurred in Japan on Saturday. Five large earthquakes occurred in central Japan on Saturday, and the epicities affected areas [...]
		Simplified Chinese	There is no obvious risk to the city and infrastructure in the past 12 hours. Numerary records for GTC were occurring at 8.4 degrees (58.8, 18.3) [...]
	joy	Spanish	The queer, introvert, positive and wonderful ever. In a happy, happy way for you. Embrace this away from the point of. Simple joy to everybody. [...]
		Simplified Chinese	Been on the monitor for an hour now, absolutely amazing work. The photo imagery, the quality of the work and the POCs are over an hour and we are pleased with our work. [...]

Table 1: Example text generations from BLOOM-560M when activating top 500 expert neurons derived from Spanish and Simplified Chinese concepts, shown at training checkpoints 10 000 and 400 000.

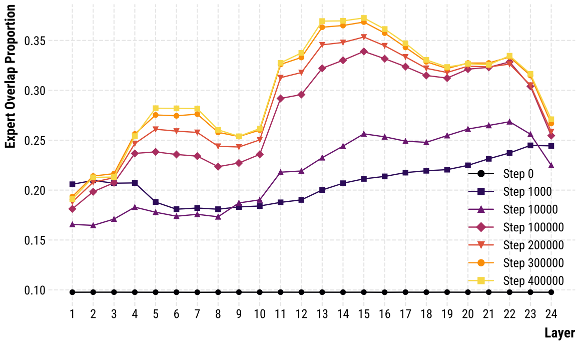


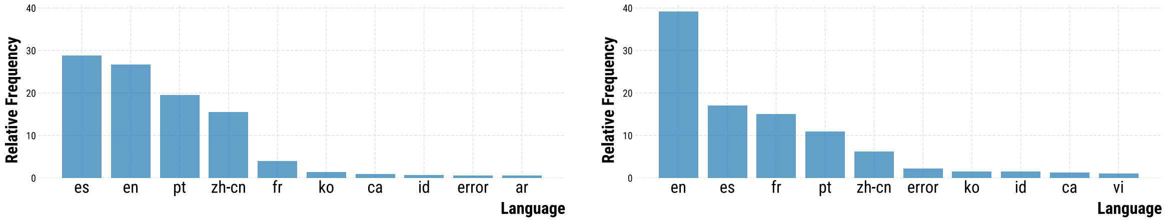
Figure 6: Cross-lingual overlap of BLOOM-560M's top 500 expert neurons per layer, showing the averaged proportion of shared neurons between language pairs.

concept (e.g., earthquake), we identify its top 500 expert neurons using data from one language (e.g., Spanish) and manipulate their activations. Specifically, we compute the median activation value of these neurons across all samples containing the concept ($b_i^{c,l} = 1$), and set their activations to these values. We then generate text by prompting the model with only a beginning-of-sequence token, using nucleus sampling ($p = 0.9$) and temperature ($t = 0.8$) across 100 random seeds. Full details are in Appendix E.

Setting neurons to their median values biases the model's representations toward the target concept. This allows us to examine whether concept-specific neurons, identified in one language, encode semantic information that generalizes across languages (or whether such neurons remain language-specific, such that neurons derived from the concept earthquake in Spanish texts lead to earthquake-related content in Spanish). Importantly, our manipulation is limited to modifying expert neurons. We provide neither language-specific nor concept-related tokens, thus giving the model freedom in choosing language and content of its generations.

Example generations are shown in Table 1. Initially, the model produces incoherent text with excessive punctuation. By step 10 000, it generates concept-relevant text in the language from which the expert neurons were derived (e.g., Spanish text about earthquakes). However, at step 400 000, while the generated text remains concept-relevant, most generations are in English, even though the manipulated expert neurons were determined using exclusively non-English data.

To quantify these observations, we analyze the language of the generated text using LANGDETECT (Shuyo, 2010). We classify 100 generations per checkpoint for all 200 concepts. For this analysis, we focus on neurons derived from Spanish data. The language distributions for steps 10 000



(a) Early training stage (step 10 000).

(b) Late training stage (step 400 000).

Figure 7: Relative frequency distribution of the top 10 detected languages when manipulating neurons derived from Spanish data, as classified using LANGDETECT.

and 400 000 are presented in Figure 7. We show distributions across all steps alongside results for neurons derived from Chinese and Swahili data in Appendix E.

At step 1000, the model is too underdeveloped to generate meaningful text, producing mainly punctuation marks that LANGDETECT fails to classify. By step 10 000, language-specific representations become most prominent–the model primarily generates Spanish text, though with a substantial presence of English. Interestingly, we also observe significant Portuguese generation, likely due to its proximity to Spanish. The notable presence of Chinese text can be attributed to its prominence in BLOOM's pre-training corpus, where it represents the second most common language after English.

This shift from generating text in the concept's source language to producing content in other languages supports our core hypothesis about the model's learning trajectory: Early training builds language-specific representations that gradually transform into compressed cross-lingual representations. Our generation experiments present direct evidence of this generalization effect in MLLMs.

While cross-lingual generalization succeeds in our experiments, its nature raises important questions. Concept knowledge successfully transfers across languages, but this transfer is biased toward high-resource languages: We observe that the model tends to express concepts in English and Chinese, regardless of the language from which these concepts were learned. This bias is particularly pronounced for neurons derived from low-resource languages like Swahili (Figure 20), which never generate in their source language. We also observe spillover within language families, such as Portuguese generation from Spanish-derived neurons. This suggests that the model uses shared neurons to form a common understanding of concepts that can be accessed across languages. However, the key question that remains is whether models can reliably draw upon such shared representations when they generate text in specific languages–especially those underrepresented in the training data.

9 Conclusion

We investigate cross-lingual generalization from a compression perspective, complementing prior and concurrent work by analyzing the pre-training process of MLLMs. Our linear probing experiments reveal a decrease in language identification performance in certain layers during pre-training, pointing to changes in how the model utilizes its parameter space. By identifying and comparing expert neurons across languages, we demonstrate that multilingual models progressively align representations across languages, ultimately sharing a substantial portion of expert neurons.

Our analysis of expert neuron distributions reveals a systematic processing pattern: The model combines token-level features from early layers with abstracted semantic content in middle layers. Notably, the proportion of shared neurons increases significantly in middle layers, indicating this is where semantic generalization primarily occurs. Generation experiments provide behavioral evidence of this phenomenon, showing the evolution from language-specific to abstracted concepts, as demonstrated by English generation from Spanish-derived concept neurons.

Future work could build on our insights to improve multilingual models. While we focused on shared representations, examining where languages maintain distinct encodings could provide supplementary understanding. Beyond obvious language-specific elements, culturally embedded concepts may require protection from the high-resource language bias we uncover. Our research offers insights for developing models that appropriately balance cross-lingual generalization with the preservation of linguistic and cultural diversity.

Limitations

Our analysis spans multiple model scales (from our small toy model to BLOOM-7B1) but does not include the largest MLLMs due to the computational demands of calculating expert neuron scores across multiple languages, concepts, and training checkpoints. Nevertheless, the observed trends appear consistent and suggest broader applicability.

We analyze the BLOOM family, which is currently the only state-of-the-art MLLM family with publicly available checkpoints. However, in both BLOOM-560M and BLOOM-7B1, some checkpoints appear to be corrupted and were excluded from our analysis. To validate our findings despite these limitations, we conduct parallel experiments with our custom toy model.

Our analysis focuses specifically on individual semantic concepts, leaving other phenomena for future work: relationships between concepts (e.g., hierarchical categories or attribute sharing), syntactic phenomena shared across languages (such as agreement and word order), and specific patterns between individual language pairs.

Ethics Statement

We do not foresee immediate ethical concerns for our research, as we primarily conduct analytical studies. BLOOM is a considerably diverse language model family with a relatively high number of underrepresented languages. While we demonstrate biases toward high-resource languages in MLLMs, potentially disadvantaging speakers of lower-resourced languages, our analysis aims to make these biases transparent. Our toy model, though potentially inheriting biases from the MC4 corpus, serves exclusively for controlled observation of MLLMs and has minimal dual-use potential.

References

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.

BigScience Workshop. 2022. BLOOM (revision 4ab0472).

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.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. 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. 2020. 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.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*.

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 generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Phys. Rev. E*, 69:066138.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. 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. 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.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Pedro Javier Ortiz Suárez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. 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.

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

Sara Rajaee and Christof Monz. 2024. Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2895–2914, St. Julian's, Malta. Association for Computational Linguistics.

Brian C. Ross. 2014. Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2):1–5.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just “OneSeC” for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709, Florence, Italy. Association for Computational Linguistics.

Nakatani Shuyo. 2010. Language detection library for java.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2022. Self-conditioning pre-trained language models. *International Conference on Machine Learning*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2022. *Natural language processing with transformers*. "O’Reilly Media, Inc.".

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. 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 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 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.

A Pre-training of Our Toy Model

We randomly initialize a model from the XGLM-564M architecture (Lin et al., 2022) and change `d_model` from 1024 to 512, resulting in a model size of approximately 257M parameters (configuration details in Table 2). We pre-train this model using the `Transformers` library (Wolf et al., 2020) with default `Trainer` parameters and a batch size of 32 for 2^{17} (= 131072) steps on the MC4 corpus (Raffel et al., 2020), which is released under the terms of ODC-BY. Pre-training language models is the intended use of this corpus. We uniformly sample data from the partitions of the following languages: `it`, `es`, `fr`, `pt`, `de`, `en`, `nl`, `af`, `zu`, `sn`, `sw`, `xh`, `ru`, `uk`, `bg`, `sr`. This sampling ensures balanced representation of all languages. We take checkpoints at powers of two $\{1, 2, 4, ..., 131072\}$ and regular 5000-step intervals. Training took 72 hours on a single NVIDIA A100-SXM4-80GB.

B Linear Probing for Language Identity

For every language that appears in both the model's pre-training data and the OSCAR (Ortiz Suárez et al., 2019) corpus, we sample 100 sentences, splitting them into 80 for training and 20 for testing. Each sentence s_i^l is tokenized into a sequence of tokens $[t_{i,0}^l, t_{i,1}^l, ..., t_{i,T_i-1}^l]$. For each tokenized sentence, the model $\mathcal{M}$ produces hidden representations: $h_{i,0}^l, h_{i,1}^l, ..., h_{i,T_1-1}^l = \mathcal{M}(t_{i,0}^l, t_{i,1}^l, ..., t_{i,T_i-1}^l)$. From these sequences of hidden states, we randomly sample one token position per sentence and extract the hidden representation at that position. We then train a logistic regression classifier for each layer to predict the language of origin for each hidden state. To ensure robustness, we repeat this experiment with three different random seeds.

As the BLOOM models have different available checkpoint intervals and our toy model is pre-trained on a different set of languages, the results are not directly comparable across models. However, the observed trends are consistent between all models regarding the evolution of representations across training checkpoints.

We present training progression results for our toy model in Figure 8 and for BLOOM-7B1 in Figure 9, with a detailed layer-wise comparison for BLOOM-7B1 shown in Figure 10.

For comparison, we include results from the encoder-only models XLM-R BASE (Figure 11) and MBERT BASE CASED (Figure 12). While decoder-only models exhibit relatively weak performance in early layers, which then increases, encoder-only models display a u-shaped pattern for language identification.

`attention_dropout`	0.1
`attention_heads`	8
`d_model`	512
`dropout`	0.1
`ffn_dim`	4096
`num_layers`	24
`vocab_size`	256008

Table 2: Configuration details of our toy model.

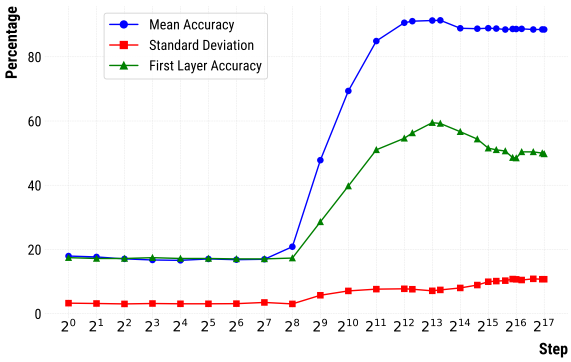


Figure 8: Language identification probing accuracy throughout training of our toy model. For each checkpoint, we show: (1) the mean accuracy across layers, (2) the standard deviation across layers (indicating how much accuracy varies between layers), and (3) the first layer's accuracy, which exhibits the most significant changes during training. Results averaged over three random seeds.

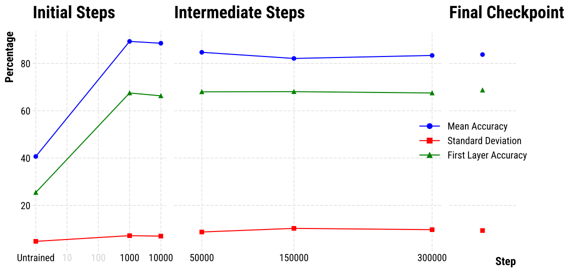
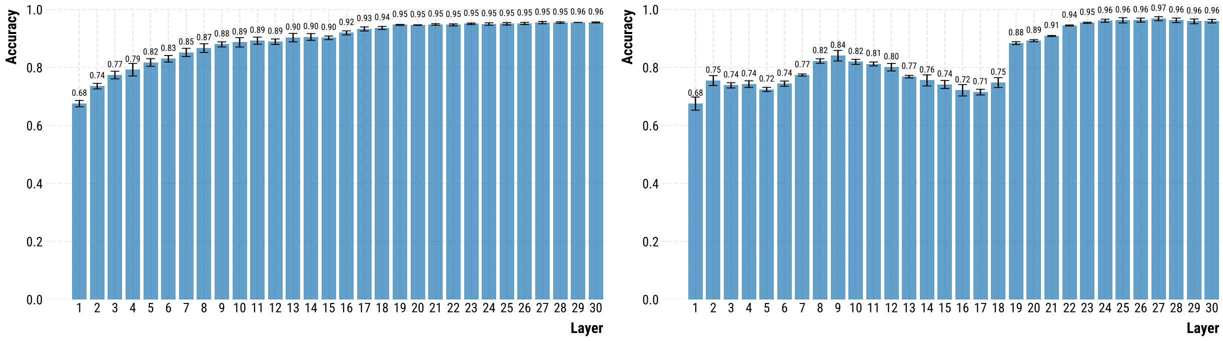


Figure 9: Language identification probing accuracy throughout training of BLOOM-7B1. For each checkpoint, we show: (1) mean accuracy across layers, (2) standard deviation across layers (indicating how much accuracy varies between layers), and (3) first layer accuracy, which exhibits the most significant changes during training. Results averaged over three random seeds.



(a) Early training stage (step 1000).

(b) Late training stage (step 300 000).

Figure 10: Language identity probing classification accuracy across layers of the BLOOM-7B1 model at different training stages. Higher accuracy indicates that language-specific information is more easily extractable from the hidden states at that layer. Error bars show standard deviation across three random seeds.

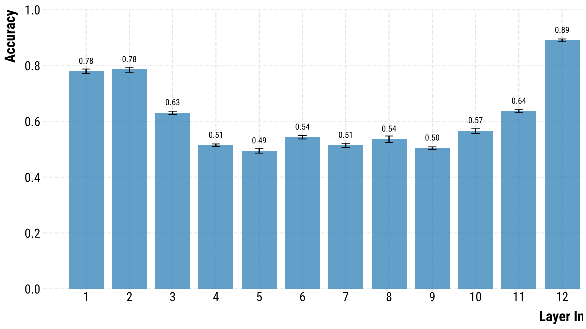


Figure 11: Layer-wise language identity probing on XLM-R BASE. Higher accuracy indicates that language-specific information is more easily extractable from the hidden states at that layer. Error bars show standard deviation across three random seeds.

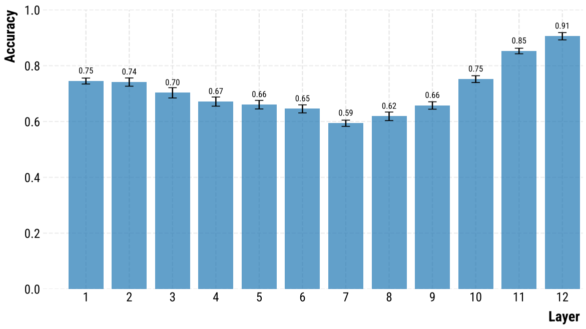


Figure 12: Layer-wise language identity probing on MBERT BASE CASED. Higher accuracy indicates that language-specific information is more easily extractable from the hidden states at that layer. Error bars show standard deviation across three random seeds.

C Expert Neuron Data

We construct binary concept identification datasets using OneSec (Scarlini et al., 2019), which provides sentences where one word per sentence is annotated with its WordNet sense. The data is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 4.0 License for research purposes. From this data, we sample 200 concepts, ensuring each has at least 100 sentences containing a word annotated with that concept. For each concept, we create negative examples by randomly sampling sentences from other concepts.

We then use NLLB 1.3B (Costa-jussà et al., 2022) with greedy decoding to translate all datasets into all of our target languages, thereby creating parallel corpora. For computational efficiency, our toy model and BLOOM-7B1 experiments use a random subset of 100 concepts. The full concept lists are available in Table 3 (BLOOM-560M) and Table 4 (BLOOM-7B1 and toy model).

D Additional Results

We show alignment matrices for all checkpoints of BLOOM-560M in Figure 13. While there is a clear increase in alignment from step 1000 to step 100 000, later checkpoints show minimal differences, becoming almost indistinguishable from each other.

Alignment matrices for the toy model are displayed in Figure 14. Early in training, a division emerges between the four Cyrillic-script languages and those using the Latin script. By step 512, while this script-based division strengthens, language families develop distinct internal alignments, visible as red squares in the matrix. Notably, script alone does not determine alignment patterns: Germanic and Romance languages (middle of matrix) show stronger mutual alignment than either does with the Bantoid languages, although all use the Latin script.

Beyond these detailed alignment analyses, we confirm that the alignment behavior during training (Figure 4) as well as the layer distributions (Figures 5 and 6) are consistently replicated in the larger BLOOM-7B1 model (Figures 15 to 17).

E Text Generation Experiments

Following Kojima et al. (2024), we generate 100 sentences per concept using different random seeds, with nucleus sampling ($p = 0.9$), temperature ($t = 0.8$), and a maximum sequence length of 64. For generation, we only manipulate the top 500 expert neurons by setting them to their concept-specific median values, and provide the model with the `</s>` token.

Example generations for the senses `earthquake-1_11_00` and `joy-1_12_00`, derived from Spanish and Simplified Chinese data, are shown in Table 1. To quantify the phenomenon of later BLOOM-560M checkpoints favoring high-resource languages, we show the detailed development for Spanish in Figure 18. Here, we see that the model first creates language-specific concepts, generating text in Spanish, but in later checkpoints favors English. Alongside English, other high-resource languages such as Chinese and French "compete for dominance" as well (step 300 000).

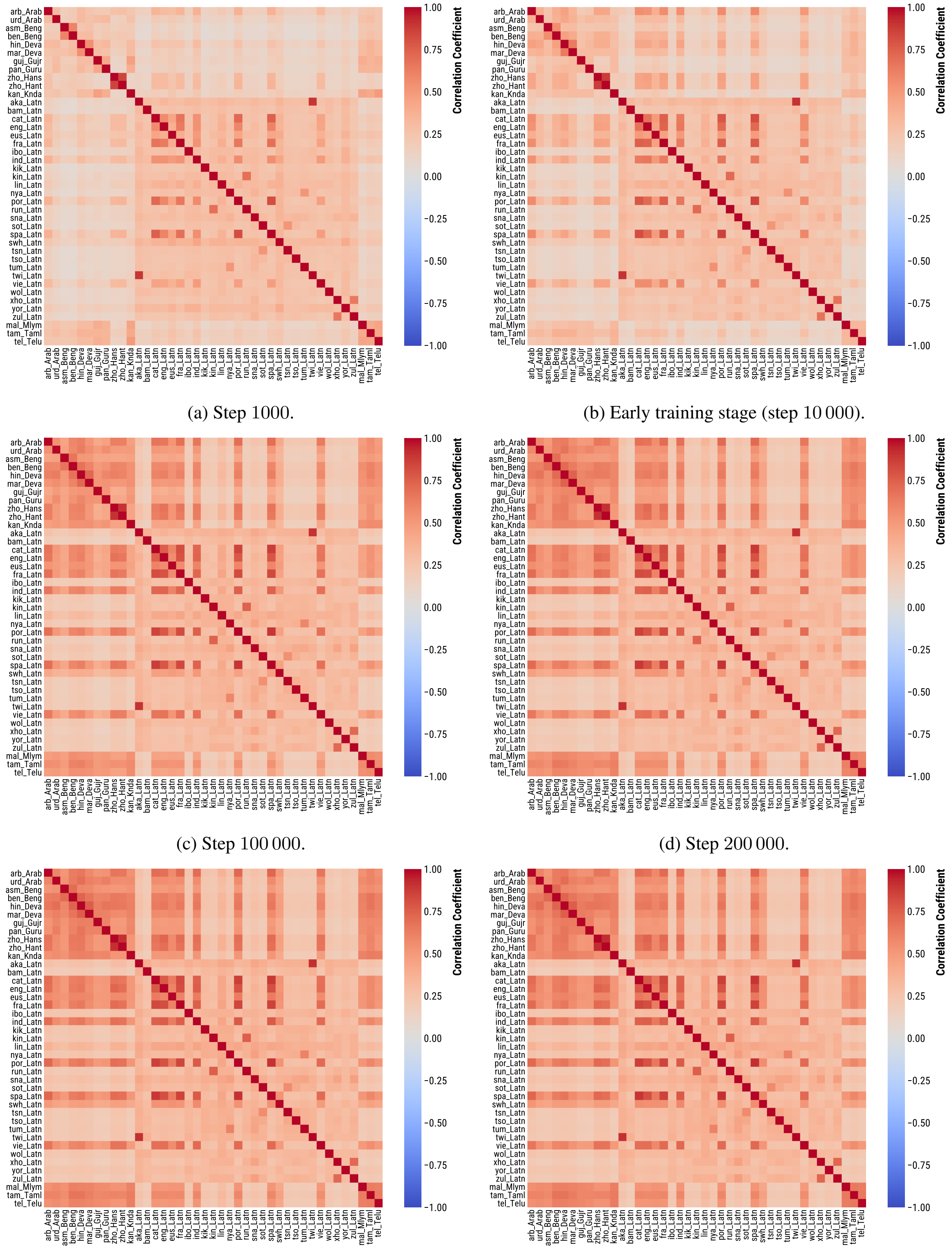
We demonstrate the same trend for Chinese in Figure 19. For lower-resourced languages like Swahili, however, the pattern differs: when deriving concept-specific neurons from Swahili data, no checkpoint generates a notable amount of Swahili text (Figure 20), suggesting these languages are underrepresented throughout training.

acceptance-1_09_00, account-1_10_00, accumulation-1_22_00, action-1_04_02, adaptation-1_22_00, adviser-1_18_00, aftershock-1_11_00, agent-1_17_00, american-1_18_00, amount-1_07_00, amount-1_21_00, appearance-1_11_00, area-1_15_01, assumption-1_09_00, attendance-1_28_00, attention-1_04_01, authority-1_18_01, backing-1_06_00, bacterium-1_05_00, band-1_14_00, bank-1_14_00, bar-1_06_04, barrel-1_06_01, bay-1_17_00, beat-1_15_00, bell-1_06_02, bill-1_10_01, body-1_14_00, boost-1_04_00, bottom-1_15_00, bourbon-1_18_01, box-1_06_02, capital-1_21_01, cent-1_21_00, center-1_15_01, ceo-1_18_00, childhood-1_26_00, church-1_06_00, circle-1_14_00, cleric-1_18_00, client-1_18_01, commitment-1_07_01, companion-1_18_02, compensation-1_22_00, conservative-1_18_00, contract-1_10_01, contractor-1_18_00, copy-1_10_00, crystal-1_27_00, cycle-1_14_00, deposit-1_19_00, desk-1_06_00, duty-1_04_00, e-mail-1_10_00, earthquake-1_11_00, economy-1_09_01, edition-1_14_00, election-1_04_01, emotion-1_12_00, end-1_15_00, enterprise-1_04_00, equity-1_21_00, equity-1_21_01, excess-1_07_02, execution-1_04_00, expulsion-1_04_01, eyebrow-1_08_00, face-1_08_00, faithful-1_14_00, family-1_14_00, favor-1_04_00, fee-1_21_00, feeling-1_03_00, find-1_04_00, forehead-1_08_00, foreigner-1_18_00, game-1_04_03, genesis-1_10_00, germany-1_15_00, goal-1_15_00, gold-1_21_00, governance-1_04_00, grievance-1_10_01, hall-1_06_03, height-1_07_00, house-1_14_01, hydrogen-1_27_00, information-1_09_00, infrastructure-1_06_00, initial-1_10_00, injection-1_27_00, insight-1_12_00, inspiration-1_06_00, interference-1_10_00, involvement-1_24_00, job-1_04_00, joy-1_12_00, judge-1_18_00, kid-1_18_00, killer-1_18_00, kind-1_09_00, kitchen-1_06_00, lack-1_26_00, lady-1_18_02, length-1_07_00, level-1_26_01, library-1_14_00, lifetime-1_28_00, machine-1_06_00, march-1_04_00, march-1_28_00, margin-1_07_00, master-1_18_00, math-1_09_00, member-1_18_00, memory-1_09_01, message-1_10_00, minister-1_18_00, ministry-1_06_00, minute-1_28_01, money-1_21_00, money-1_21_02, morale-1_07_00, move-1_04_01, mr-1_10_00, mystery-1_09_00, need-1_17_00, negotiation-1_10_00, news-1_10_01, nickname-1_10_01, nobility-1_14_00, notion-1_09_00, one-1_23_00, order-1_07_01, paint-1_06_00, paradox-1_10_00, participant-1_18_00, participant-1_18_01, pattern-1_09_00, percent-1_24_00, percentage-1_24_00, perimeter-1_25_00, person-1_03_00, personality-1_18_00, pet-1_05_00, phase-1_26_00, phosphorus-1_27_00, pier-1_06_00, place-1_15_04, politician-1_18_01, poster-1_18_00, premonition-1_12_00, president-1_18_01, president-1_18_04, process-1_04_00, program-1_09_00, programme-1_10_00, pub-1_06_00, race-1_11_00, rank-1_14_00, recovery-1_11_00, refinery-1_06_00, regard-1_09_01, release-1_04_01, release-1_06_00, role-1_04_00, schoolteacher-1_18_00, score-1_10_00, scourge-1_26_00, senator-1_18_00, september-1_28_00, signal-1_16_00, situation-1_26_01, skepticism-1_09_01, solution-1_27_00, someone-1_03_00, space-1_03_00, spain-1_15_00, spite-1_12_00, statement-1_10_00, step-1_04_02, striker-1_18_02, suicide-1_04_00, suspension-1_28_00, system-1_06_00, tax-1_21_00, thing-1_04_00, thinking-1_09_00, times-1_04_00, triage-1_04_00, trial-1_04_00, type-1_18_00, unemployment-1_26_00, verdict-1_04_00, vicar-1_18_00, wealth-1_26_00, wednesday-1_28_00, will-1_09_00, yield-1_04_00, zip-1_10_00

Table 3: Complete set of 200 randomly sampled WORDNET senses (alphabetically ordered) used in BLOOM-560M experiments.

accumulation-1_22_00, action-1_04_02, adviser-1_18_00, aftershock-1_11_00, american-1_18_00, amount-1_07_00, appearance-1_11_00, attention-1_04_01, authority-1_18_01, bar-1_06_04, boost-1_04_00, bourbon-1_18_01, center-1_15_01, ceo-1_18_00, circle-1_14_00, client-1_18_01, companion-1_18_02, conservative-1_18_00, contractor-1_18_00, cycle-1_14_00, deposit-1_19_00, e-mail-1_10_00, economy-1_09_01, edition-1_14_00, equity-1_21_00, excess-1_07_02, execution-1_04_00, eyebrow-1_08_00, faithful-1_14_00, family-1_14_00, favor-1_04_00, forehead-1_08_00, foreigner-1_18_00, genesis-1_10_00, germany-1_15_00, goal-1_15_00, gold-1_21_00, governance-1_04_00, height-1_07_00, house-1_14_01, information-1_09_00, infrastructure-1_06_00, insight-1_12_00, inspiration-1_06_00, interference-1_10_00, job-1_04_00, joy-1_12_00, kid-1_18_00, killer-1_18_00, lady-1_18_02, length-1_07_00, library-1_14_00, lifetime-1_28_00, march-1_04_00, margin-1_07_00, master-1_18_00, message-1_10_00, money-1_21_00, morale-1_07_00, move-1_04_01, mystery-1_09_00, negotiation-1_10_00, news-1_10_01, nobility-1_14_00, notion-1_09_00, paint-1_06_00, participant-1_18_00, participant-1_18_01, pattern-1_09_00, percent-1_24_00, perimeter-1_25_00, personality-1_18_00, pet-1_05_00, phosphorus-1_27_00, pier-1_06_00, place-1_15_04, premonition-1_12_00, president-1_18_01, president-1_18_04, program-1_09_00, pub-1_06_00, race-1_11_00, rank-1_14_00, refinery-1_06_00, release-1_06_00, september-1_28_00, skepticism-1_09_01, someone-1_03_00, spite-1_12_00, striker-1_18_02, system-1_06_00, tax-1_21_00, thing-1_04_00, thinking-1_09_00, type-1_18_00, unemployment-1_26_00, verdict-1_04_00, vicar-1_18_00, wealth-1_26_00, yield-1_04_00

Table 4: Subset of 100 WORDNET senses randomly selected from Table 3 (alphabetically ordered), used in BLOOM-7B1 and toy model experiments for faster experimentation.



(a) Step 1000.

(b) Early training stage (step 10 000).

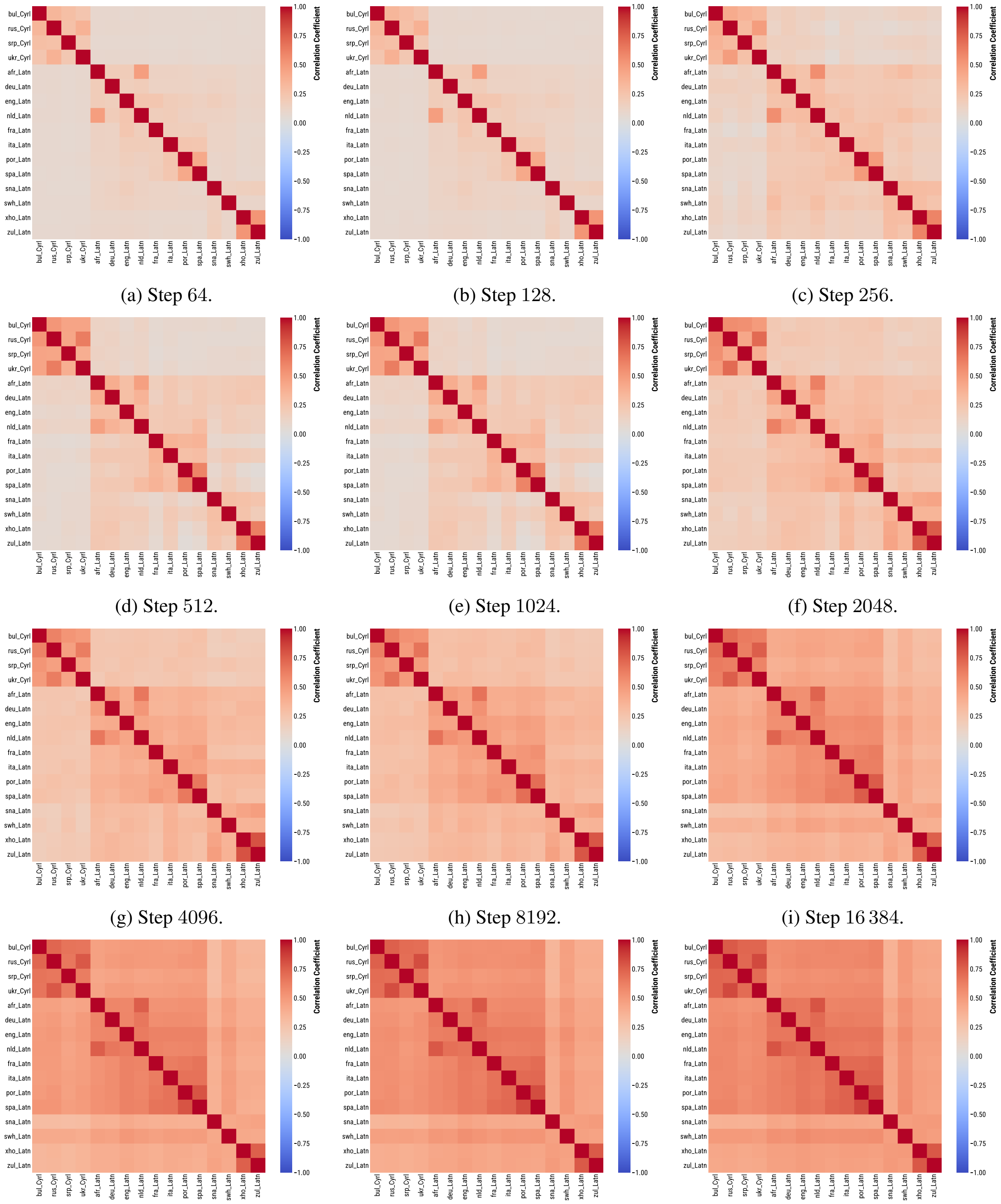
(c) Step 100 000.

(d) Step 200 000.

(e) Step 300 000.

(f) Step 400 000.

Figure 13: Expert neuron alignment of BLOOM-560M at different training stages.



(a) Step 64.

(b) Step 128.

(c) Step 256.

(d) Step 512.

(e) Step 1024.

(f) Step 2048.

(g) Step 4096.

(h) Step 8192.

(i) Step 16 384.

(j) Step 32 768.

(k) Step 65 536.

(l) Step 131 072.

Figure 14: Expert neuron alignment of our toy model at different training stages.

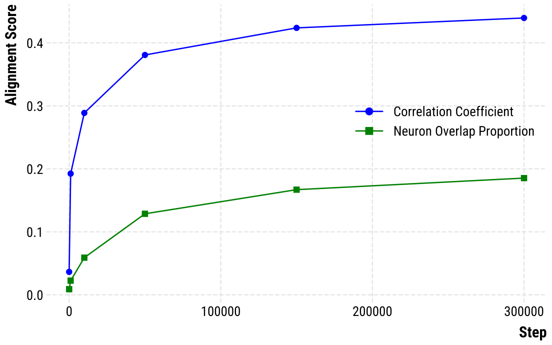


Figure 15: Correlation Coefficient and Neuron Overlap Proportion of top 500 neurons throughout training, averaged across concepts and languages for BLOOM-7B1.

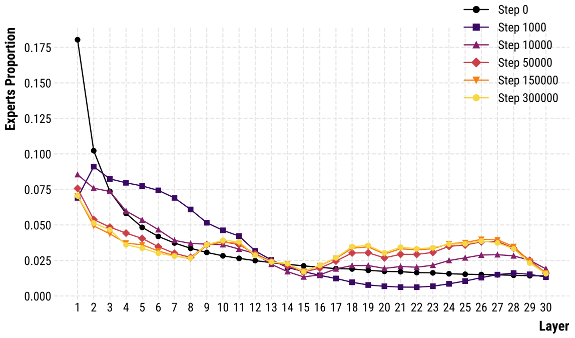


Figure 16: Layer-wise distribution of BLOOM-7B1's top 500 expert neurons, averaged across languages and concepts.

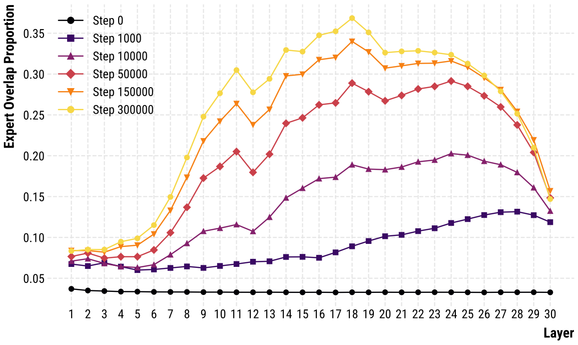
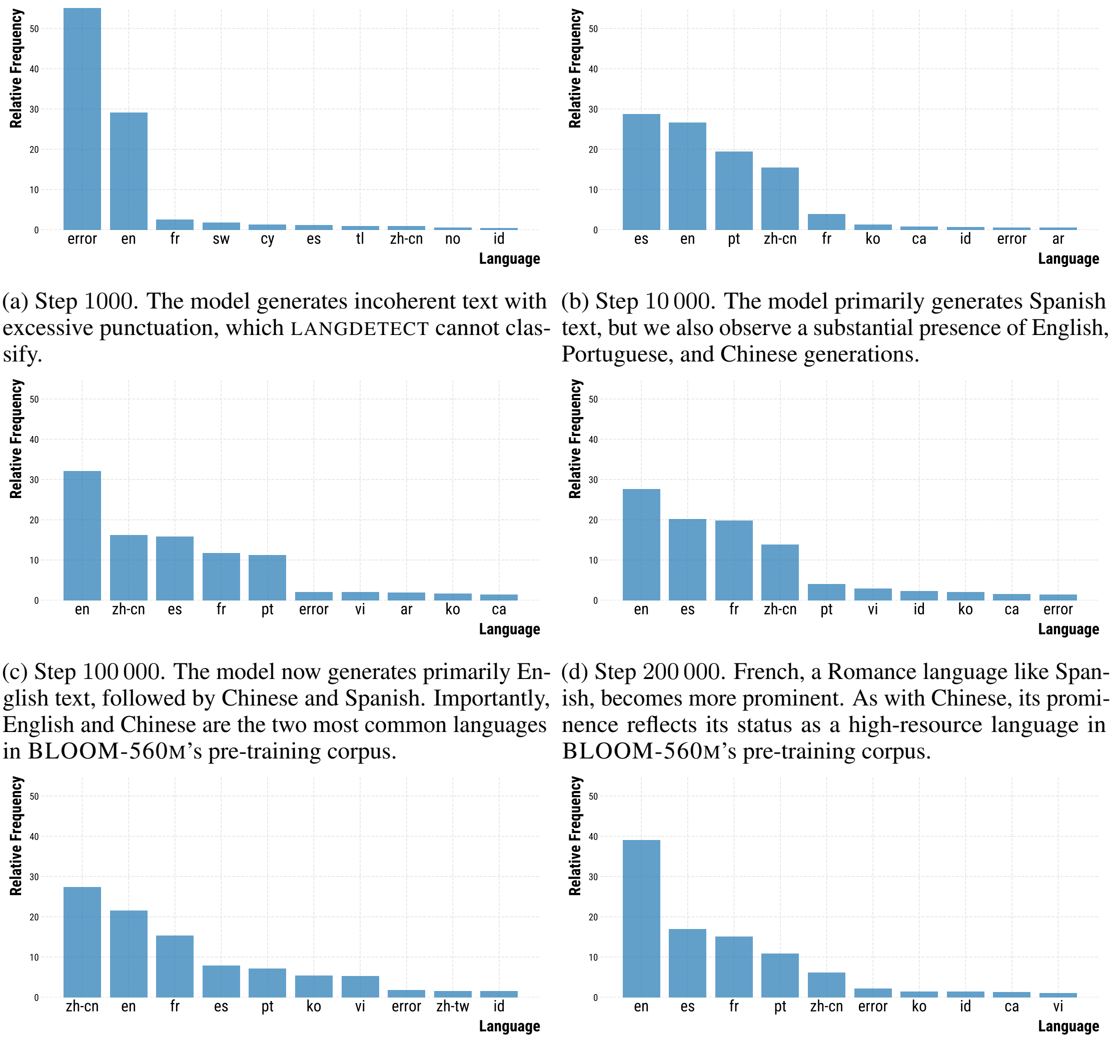


Figure 17: Cross-lingual overlap of BLOOM-7B1's top 500 expert neurons per layer, showing the averaged proportion of shared neurons between language pairs.



(a) Step 1000. The model generates incoherent text with excessive punctuation, which LANGDETECT cannot classify.

(b) Step 10 000. The model primarily generates Spanish text, but we also observe a substantial presence of English, Portuguese, and Chinese generations.

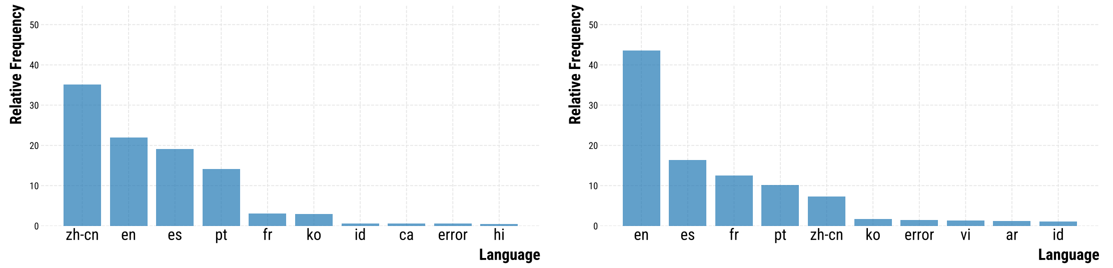
(c) Step 100 000. The model now generates primarily English text, followed by Chinese and Spanish. Importantly, English and Chinese are the two most common languages in BLOOM-560M's pre-training corpus.

(d) Step 200 000. French, a Romance language like Spanish, becomes more prominent. As with Chinese, its prominence reflects its status as a high-resource language in BLOOM-560M's pre-training corpus.

(e) Step 300 000. The three high-resource languages–Chinese, English, and French–dominate, while Spanish becomes increasingly less present.

(f) Step 400 000. The model generates a large amount of English text, with Spanish appearing to a much lesser extent.

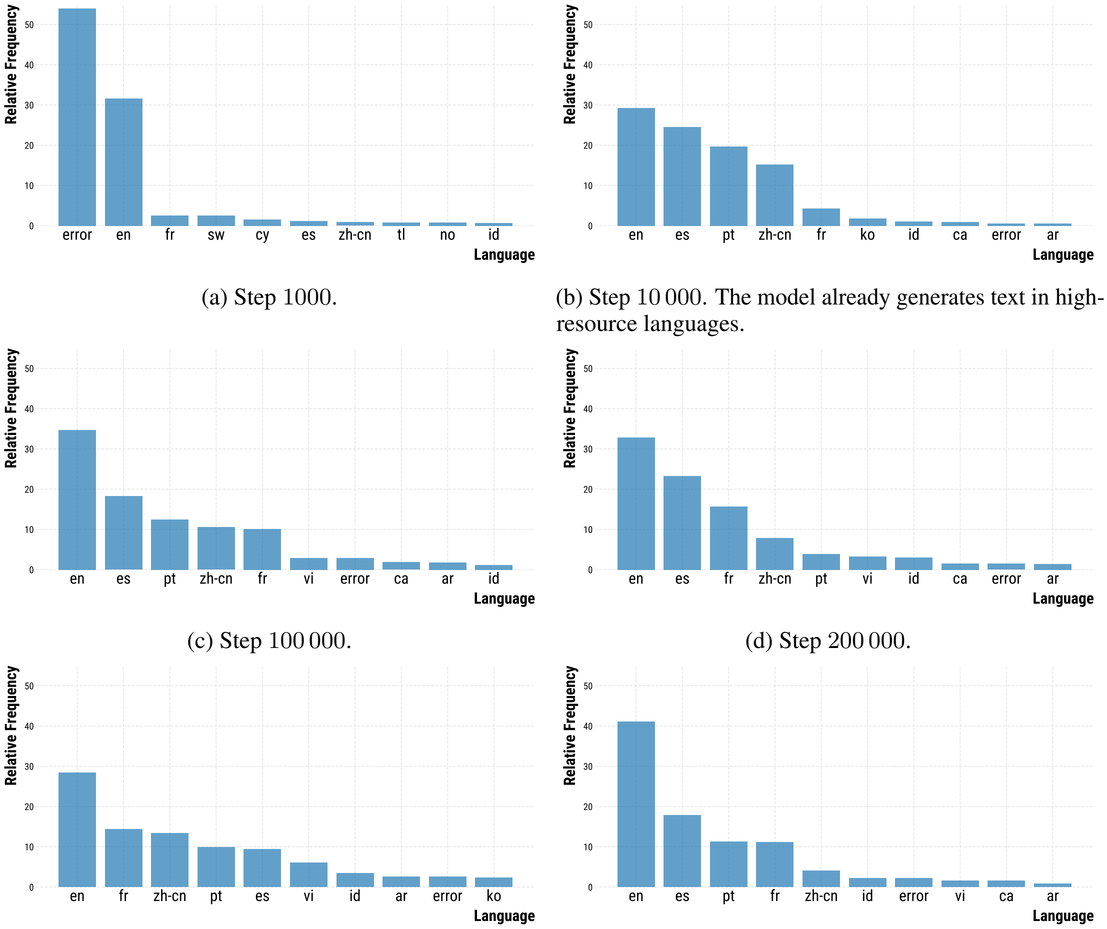
Figure 18: Text generation experiments: Relative frequency distribution of the top 10 detected languages when manipulating neurons derived from Spanish data across all available BLOOM-560M checkpoints, as classified using LANGDETECT.



(a) Early training stage (step 10 000).

(b) Late training stage (step 400 000).

Figure 19: Text generation experiments: Relative frequency distribution of the top 10 detected languages when manipulating neurons derived from Chinese data, as classified using LANGDETECT.



(a) Step 1000.

(b) Step 10 000. The model already generates text in high-resource languages.

(c) Step 100 000.

(d) Step 200 000.

(e) Step 300 000.

(f) Step 400 000.

Figure 20: Text generation experiments: Relative frequency distribution of the top 10 detected languages when manipulating neurons derived from Swahili data across all available BLOOM-560M checkpoints, as classified using LANGDETECT.