

# How Does Continued African Adaptation Reshape Multilingual Representations?

Oluwadarasimi Olowookere-Ajayi and Jesse-Paul Osemeke  
ML Collective

## Abstract

Continued pretraining has become a widely adopted approach for improving multilingual large language models, consistently improving performance across a diverse range of downstream NLP tasks. While these gains are well documented, little is known about how continued pretraining reshapes the internal representations that give rise to these improvements. In this work, we investigate how African-language continued pretraining reorganizes the internal representation space of multilingual language models. We compare three pretrained multilingual models, XLM-R, Gemma-3, and Llama-3.1, with their African-adapted counterparts, AfroXLMR, AfriqueGemma, and AfriqueLlama, spanning both encoder-only and decoder-only architectures. Using layer-wise representation similarity, representation geometry, linear probing, and cross-lingual transfer analyses, we characterize how continued pretraining reorganizes multilingual representations across encoder-only and decoder-only architectures. Our analysis shows that continued African-language pretraining selectively reorganizes multilingual representations rather than uniformly modifying them. The patterns of reorganization differ across the model pairs studied. Continued pretraining expands the effective dimensionality of the representation space. For the encoder pair, adaptation also improves POS accessibility and cross-lingual transfer across African languages. Together, these findings provide new insights into how continued pretraining reshapes multilingual representation spaces and advances our understanding of multilingual adaptation for underrepresented languages.

## 1 Introduction

Multilingual pretrained language models such as XLM-R (Conneau et al., 2020a) and Gemma 3 (Gemma Team, 2025) have significantly advanced multilingual NLP. These models learn

shared multilingual representations that facilitate cross-lingual transfer, allowing knowledge acquired from high-resource languages to benefit lower-resource ones. However, due to the trade-offs of model capacity and uneven web data distribution (often referred to as the curse of multilinguality (Conneau et al., 2020a)), many low-resource languages, particularly African languages, remain severely underrepresented during initial pre-training (Adebara et al., 2025). To address this gap without having to train massive models from scratch, continued pretraining (CPT) on African language corpora has emerged as an effective adaptation strategy, leading to models such as AfroXLMR (Alabi et al., 2022) and AfriqueLLM (Yu et al., 2026) that consistently outperform their multilingual counterparts (Gemma Team (2025), Conneau et al. (2020a)) across a range of African NLP tasks.

While these results demonstrate that continued pretraining is effective, they provide limited insight into how the adaptation process reshapes the model’s internal representations. As a result, it remains unclear what linguistic knowledge is acquired during continued pretraining, which layers of the network undergo the greatest change, and how these representational changes contribute to improved downstream performance. Developing a deeper understanding of the internal effects of continued pretraining is essential for designing effective multilingual adaptation strategies and for explaining why adapted models succeed where their base counterparts do not.

To investigate how continued African-language pretraining reshapes multilingual representations, we address the following research questions:

1. **RQ1:** How does continued African-language pretraining reshape the internal representations of multilingual language models?
2. **RQ2:** How does continued African-language

pretraining affect the accessibility and cross-lingual transfer of linguistic knowledge across transformer layers?

3. **RQ3:** How does continued African-language pretraining reshape the geometry of multilingual representation spaces?

To answer these questions, we conduct a comprehensive representation analysis across three pairs of multilingual language models before and after African-language continued pretraining and analyze representation evolution from four perspectives. Our analysis reveals that continued African-language pretraining selectively reorganizes multilingual representations rather than uniformly modifying the network.

## 1.1 Our Contributions

Our contributions are summarized as follows:

- We present a systematic analysis of representation evolution during African-language continued pretraining across both encoder-only and decoder-only multilingual language models, directly comparing pretrained models with their African-adapted counterparts.
- Our joint analysis shows that representational drift and dimensional expansion need not coincide: in AfroXLMR, their peaks occur at different layers, and some layers drift while losing effective rank. For the XLM-R-large/AfroXLMR-large encoder pair, POS probing further reveals improved cross-lingual transfer and a shift in its peak toward intermediate layers.

## 2 Related Work

### 2.1 African Multilingual Language Models and Continued Pretraining

The underrepresentation of African languages in large-scale pretraining corpora has motivated the development of African-focused multilingual models. AfriBERTa showed that competitive multilingual models could be trained on less than one gigabyte of text covering eleven African languages (Ogueji et al., 2021); subsequent efforts expanded coverage through models such as AfroLM, pretrained from scratch for 23 African languages using self-active learning (Dossou et al., 2022), and Serengeti, which covers hundreds of African

languages and language varieties (Adebara et al., 2023).

Continued pretraining offers an alternative to training new models from scratch. Gururangan et al. (2020) demonstrate that domain- and task-adaptive pretraining improves text classification across domains. For African languages, AfroXLMR continues masked-language-model pretraining from XLM-R and improves both supervised and zero-shot cross-lingual performance (Alabi et al., 2022), while AfriqueLLM extends continued pretraining to decoder-only families including Gemma, Llama, and Qwen across 20 African languages (Yu et al., 2026). The closest representation-level study, Aduah and Meyer (2025), uses layer-wise probing to show that African-focused models encode more recoverable linguistic information for their target languages. We extend this work by comparing representation drift and geometry across base-adapted encoder and decoder pairs, with POS probing and cross-lingual transfer evaluation on the encoder pair.

### 2.2 Geometric and Layer-wise Representation Analysis

Understanding the internal representations learned by deep neural networks has become a central problem in model interpretability. Early studies showed that linguistic information is distributed unevenly across transformer layers, with lower layers capturing lexical features, intermediate layers encoding syntactic information, and upper layers becoming increasingly contextual and task-specific (Jawahar et al., 2019). Merchant et al. (2020) find that supervised fine-tuning changes BERT representations primarily in upper layers, with task-dependent variation; our study examines continued pretraining rather than supervised fine-tuning. To compare representations directly, canonical-correlation methods such as SVCCA were introduced (Raghu et al., 2017; Morcos et al., 2018), followed by Centered Kernel Alignment (CKA), which has become a standard tool for measuring layer-wise representation similarity across independently trained models because of its robustness and invariance properties (Kornblith et al., 2019). These methods have been widely used to study training dynamics, representation evolution, and layer correspondence in large language models (Chiang et al., 2020).

Recent work has extended representation analysis beyond similarity to examine the geom-

etry of representation spaces. Studies have shown that contextual embeddings are often highly anisotropic, occupying only a small number of dominant directions (Ethayarajh, 2019), and that encoder and decoder models exhibit distinct geometric properties across layers (Razzhigaev et al., 2024). Similar geometric analyses have also been applied to multilingual machine translation, revealing systematic differences in the intrinsic dimensionality and isotropy of multilingual representations (Verma et al., 2024). Despite these advances, little is known about how continued pretraining reshapes the geometry and layer-wise organization of multilingual representations for African languages. Our work addresses this gap by combining CKA with geometric analyses to characterize how continued African-language pretraining reorganizes multilingual representation spaces.

### 2.3 Cross-lingual Alignment and Knowledge Transfer

Multilingual language models learn representations that enable knowledge acquired from one language to transfer to others with little or no task-specific supervision. Early studies of multilingual BERT demonstrated strong zero-shot transfer across languages, although performance generally decreases as typological and lexical differences increase (Pires et al., 2019; Wu and Dredze, 2019). Subsequent work attributed this ability to shared multilingual representations emerging from joint pretraining, parameter sharing, lexical overlap, and typological similarity, while showing that effective transfer can also arise in the absence of shared vocabularies (Conneau et al., 2020b; Artetxe et al., 2020; Deshpande et al., 2022). Together, these studies established cross-lingual representation learning as the primary mechanism underlying multilingual transfer.

To understand what multilingual representations encode, prior work has extensively employed diagnostic probing, where lightweight classifiers are trained on frozen representations to measure the accessibility of linguistic information (Alain and Bengio, 2017; Hewitt and Liang, 2019). Layer-wise probing has shown that linguistic information is distributed unevenly throughout transformer networks, with different layers specializing in lexical, syntactic, and semantic knowledge (Jawahar et al., 2019; Fayyaz et al., 2021). More recent work has extended probing to African multilingual models, showing that

syntactic information is generally most accessible in the middle and upper transformer layers (Aduah and Meyer, 2025). However, existing studies primarily analyze pretrained models in isolation. It remains unclear how continued African-language pretraining redistributes linguistic knowledge across transformer layers or how these representational changes influence cross-lingual transfer. Our work addresses this gap by combining layer-wise probing with cross-lingual transfer evaluation to characterize how continued pretraining reshapes multilingual knowledge sharing.

## 3 Methodology

Let  $M_b$  denote a pretrained multilingual language model and  $M_a$  its counterpart after continued African-language pretraining. For an input sequence  $X$ , we denote the hidden representation produced at layer  $l$  by  $H^{(l)} \in \mathbb{R}^{n \times d}$ . We characterize the representational changes induced by continued pretraining through three complementary analyses: representation similarity, representation geometry, and linguistic probing with cross-lingual transfer.

### 3.1 Representation Similarity Analysis

Let  $M_b$  denote a pretrained multilingual language model and  $M_a$  its African-adapted counterpart obtained through continued pretraining. To quantify how continued pretraining changes the internal representations of multilingual language models, we measure the similarity between the hidden representations of a pretrained model and its African-adapted counterpart using Centered Kernel Alignment (CKA) (Kornblith et al., 2019). CKA provides a robust similarity measure that is invariant to isotropic scaling and orthogonal transformations, making it well suited for comparing representations learned by independently trained neural networks.

For each transformer layer  $l$ , let

$$Z_b^{(l)}, Z_a^{(l)} \in \mathbb{R}^{N \times d}$$

denote the sentence representation matrices extracted from the pretrained model  $M_b$  and its African-adapted counterpart  $M_a$ , respectively, where  $N$  is the number of input sentences and  $d$  is the hidden dimension.

Before computing CKA, we center each representation dimension across sentences.

Following Kornblith et al. (2019), we compute the linear Centered Kernel Alignment (CKA) between the corresponding layer representations as

$$\text{CKA}(Z_b^{(l)}, Z_a^{(l)}) = \frac{\left\| \left( Z_a^{(l)} \right)^\top Z_b^{(l)} \right\|_F^2}{\left\| \left( Z_b^{(l)} \right)^\top Z_b^{(l)} \right\|_F \left\| \left( Z_a^{(l)} \right)^\top Z_a^{(l)} \right\|_F} \times \frac{1}{\left\| \left( Z_a^{(l)} \right)^\top Z_a^{(l)} \right\|_F}.$$

where  $\| \cdot \|_F$  denotes the Frobenius norm. We compute CKA independently for every corresponding transformer layer to obtain a layer-wise similarity profile between the pretrained and adapted models. Lower CKA values indicate greater representational change following continued pretraining, while higher values suggest that the learned representations remain largely preserved.

### 3.2 Representation Geometry Analysis

CKA quantifies how much representations change but does not characterize the geometry of the resulting representation space. We therefore analyze the effective dimensionality and spectral structure of the sentence representation matrix  $Z^{(l)} \in \mathbb{R}^{N \times d}$ .

Let

$$Z = U \Sigma V^\top$$

be its singular value decomposition, with singular values  $\{\sigma_i\}_{i=1}^r$ . Following Roy and Vetterli (2007), we define

$$p_i = \frac{\sigma_i}{\sum_{j=1}^r \sigma_j},$$

$$\text{erank}(Z) = \exp \left( - \sum_{i=1}^r p_i \log p_i \right).$$

Higher effective rank indicates that variance is distributed across more latent directions. We additionally examine the singular-value spectrum, principal-component variance, anisotropy, and local neighbourhood preservation to characterize complementary aspects of representation geometry.

### 3.3 Linguistic Knowledge and Cross-lingual Transfer Analysis

To investigate how continued pretraining influences multilingual linguistic representations, we

employ linear probing (Alain and Bengio, 2017) on token-level hidden representations. Linear probes provide a simple yet effective mechanism for measuring the accessibility of linguistic information while minimizing the expressive capacity of the classifier (Hewitt and Liang, 2019). We use the same probing framework to evaluate two complementary properties of the learned representations:

- the layer-wise accessibility of part-of-speech information and
- the ability of linguistic knowledge learned from one language to transfer to another.

## 4 Experimental Setup

Unless otherwise stated, model parameters remain frozen during representation extraction and probe training.

### 4.1 Models

We compare three pretrained models with their African-adapted counterparts: XLM-R-large with AfroXLMR-large, Gemma-3-4B with AfriqueGemma-4B, and Llama-3.1-8B with AfriqueLlama-8B. The first pair is encoder-only, while the latter two are decoder-only. Model configurations are summarized in Table 1.

XLM-R-large and AfroXLMR-large follow the masked-language-modeling objective, with AfroXLMR obtained by continued pretraining on African-language corpora (Alabi et al., 2022). AfriqueGemma-4B and AfriqueLlama-8B were similarly obtained by continuing the pretraining of Gemma-3-4B and Llama-3.1-8B on approximately 26B multilingual tokens, including 20 African languages (Yu et al., 2026). Since the adapted checkpoints preserve the tokenizer and architecture of their corresponding base models, identical tokenized inputs can be supplied when comparing representations.

### 4.2 Datasets

We use two datasets corresponding to the sentence-level and token-level analyses performed in this work. FLEURS (Conneau et al., 2023) provides multilingual sentence transcriptions for representation similarity and geometry analyses, while MasakhaPOS (Dione et al., 2023) provides token-level part-of-speech annotations for probing and cross-lingual transfer experiments.

| Model Pair                     | Architecture | Parameters | Layers | Hidden Dim. | Objective |
|--------------------------------|--------------|------------|--------|-------------|-----------|
| XLM-R-large → AfroXLMR-large   | Encoder      | 560M       | 24     | 1024        | MLM       |
| Gemma-3-4B → AfriqueGemma-4B   | Decoder      | 4B         | 34     | 2560        | Causal LM |
| Llama-3.1-8B → AfriqueLlama-8B | Decoder      | 8B         | 32     | 4096        | Causal LM |

Table 1: Model pairs used throughout this work.

### 4.3 Hidden-State Representation Extraction

We extract the embedding output and every transformer-layer hidden state with model parameters frozen. For FLEURS, sentence representations are obtained at each layer by mean pooling over non-padding and non-special tokens and are used for CKA and geometry analyses. For MasakhaPOS, we retain token-level representations, assigning each POS label to the first subword and excluding subsequent subwords, padding, and special tokens.

### 4.4 Probe Training

We train an independent linear probe for every hidden-state level of XLM-R-large and AfroXLMR-large. Each probe consists of a single affine projection from the hidden dimension to the 17 Universal POS tags without additional hidden layers or nonlinearities.

All probes are optimized using AdamW with a learning rate of  $10^{-3}$  for 30 epochs with early stopping. The pretrained language model remains frozen throughout training, ensuring that probe performance reflects the linear accessibility of linguistic information rather than additional representation learning.

For cross-lingual transfer, a probe trained on one source language is directly evaluated on every target language without further adaptation, producing a complete source–target transfer matrix for every transformer layer.

### 4.5 Evaluation Metrics

**Representation Geometry.** We characterize the geometry of each sentence representation matrix using effective rank, participation ratio, singular value spectra, and principal-component variance. We additionally estimate anisotropy using the average cosine similarity between randomly sampled sentence pairs and analyze local neighborhood preservation before and after continued pre-training.

**Linguistic Knowledge.** We evaluate the accessibility of linguistic information using token-level

POS probing. Performance is measured using macro-averaged F1 and accuracy, with Macro-F1 serving as the primary metric because it equally weights all POS categories regardless of class frequency.

**Cross-lingual Transfer.** Cross-lingual transfer is measured by evaluating probes trained on a source language directly on target-language test sets without additional supervision. Transfer performance is reported using Macro-F1 and accuracy. For selection-free aggregate comparisons, we first average performance across all 25 hidden-state levels for each source–target pair and then average diagonal and off-diagonal pairs separately. Best-layer results are reported only as post hoc descriptive analyses in the appendix.

## 5 Results

### 5.1 RQ1: How Does Continued Pretraining Reshape Internal Representations?

Figure 1 shows layer-wise representational plasticity, defined as  $1 - \text{CKA}$ , between each pre-trained model and its African-adapted counterpart. Across all three model families, the embedding representations remain largely unchanged, while substantially greater differences emerge within the transformer layers. The location of these changes, however, differs markedly across architectures.

For XLM-R-large, mean CKA is 0.986 at the embedding layer and decreases progressively with depth, reaching its minimum of 0.218 at layer 23 before partially recovering in the final layer. AfroXLMR therefore exhibits its strongest representational change in the upper encoder layers. The magnitude of this change also varies across languages: Hausa, Wolof, and Igbo show the largest average drift, whereas Yoruba changes substantially less. English and French also exhibit non-negligible drift, indicating that adaptation affects the broader multilingual representation space rather than only the target African languages.

AfriqueGemma follows a different depth profile. Its embedding representation is almost unchanged ( $\text{CKA} = 0.999$ ), but similarity decreases rapidly in

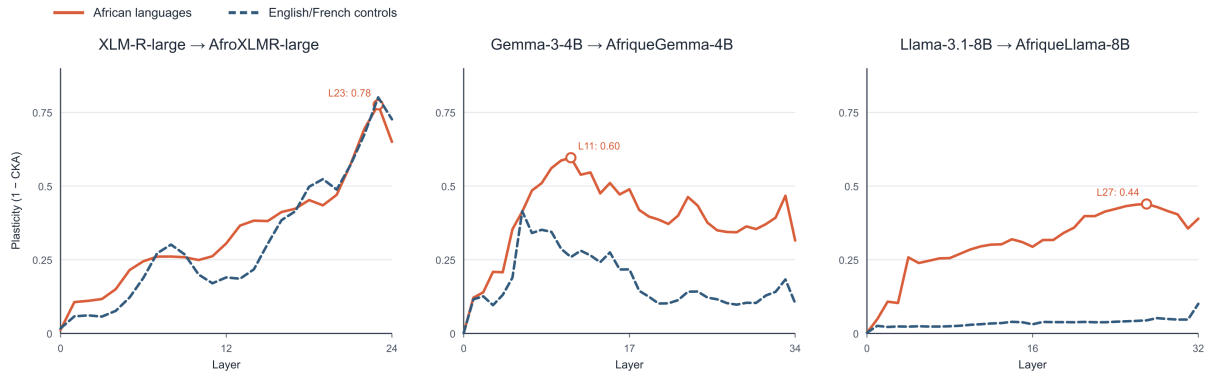


Figure 1: Layer-wise sentence-representation plasticity, measured as  $1 - \text{CKA}$ , for each pretrained–adapted model pair. Curves average six African languages (solid) and the English/French controls (dashed). Embedding representations remain stable, but the depth of maximum change differs sharply across architectures: upper layers for XLM-R, early-to-middle layers for Gemma, and later layers for Llama.

the early transformer blocks and reaches its minimum around layers 10–11 ( $\text{CKA} \approx 0.488$ ) before increasing again toward the upper layers. Representational change is also more selective by language: average drift across the African languages is more than twice that observed for English and French.

AfriqueLlama shows a third pattern: change accumulates more gradually across depth and is strongly language selective. Mean African-language plasticity rises from 0.002 at the embedding layer to 0.439 at layer 27 and remains high at the final layer (0.389). In contrast, the English/French control curve remains below 0.10 until the final layer. Continued pretraining therefore preserves the input embedding space in all three architectures but reorganizes contextual representations at architecture-specific depths: predominantly in the upper layers of XLM-R, the early-to-middle layers of Gemma, and the later layers of Llama.

Language-specific plasticity profiles for all three architectures are reported in Appendix Figure 4.

## 5.2 RQ2: How Does Continued Pretraining Affect Cross-lingual Knowledge Transfer?

Figure 2 shows that continued pretraining yields substantially larger gains in cross-lingual transfer than in in-language POS probing. To avoid selecting layers using test performance, we average each source–target pair over all 25 hidden-state levels. Cross-lingual macro-F1 increases from 17.3% for XLM-R-large to 24.1% for AfroXLMR-

large, a gain of 6.8 points, while cross-lingual accuracy increases from 35.0% to 43.2%. In comparison, in-language macro-F1 increases from 64.9% to 67.6%, a gain of 2.8 points, and accuracy increases from 78.8% to 81.9%. The gap between in-language and cross-lingual macro-F1 therefore narrows by approximately four points, indicating that adaptation has a larger effect on transferability than on in-language POS accessibility.

Continued pretraining also changes where transferable information is most accessible. Mean cross-lingual macro-F1 peaks at layer 2 in XLM-R-large but shifts to layer 12 in AfroXLMR-large. Cross-lingual accuracy shows a similar shift, from layer 10 to layer 11, while AfroXLMR also retains stronger transfer performance in its upper layers. The improvement is therefore not a uniform increase across depth; transferable POS information becomes more accessible in the intermediate layers after adaptation.

Transfer gains remain uneven across languages. When performance is averaged across all hidden-state levels, Igbo and Yoruba benefit most as targets, with average incoming cross-lingual macro-F1 gains of 11.6 and 10.0 points, respectively. Hausa and Swahili improve by 6.0 and 4.9 points, whereas Wolof improves by only 1.5 points. Four individual transfer directions show modest decreases, none exceeding 2.1 points (Figure 2).

Appendix Figure 5 reports the post hoc best-layer transfer matrices, while Appendix Table 2 summarizes the selection-free layer-averaged scores.

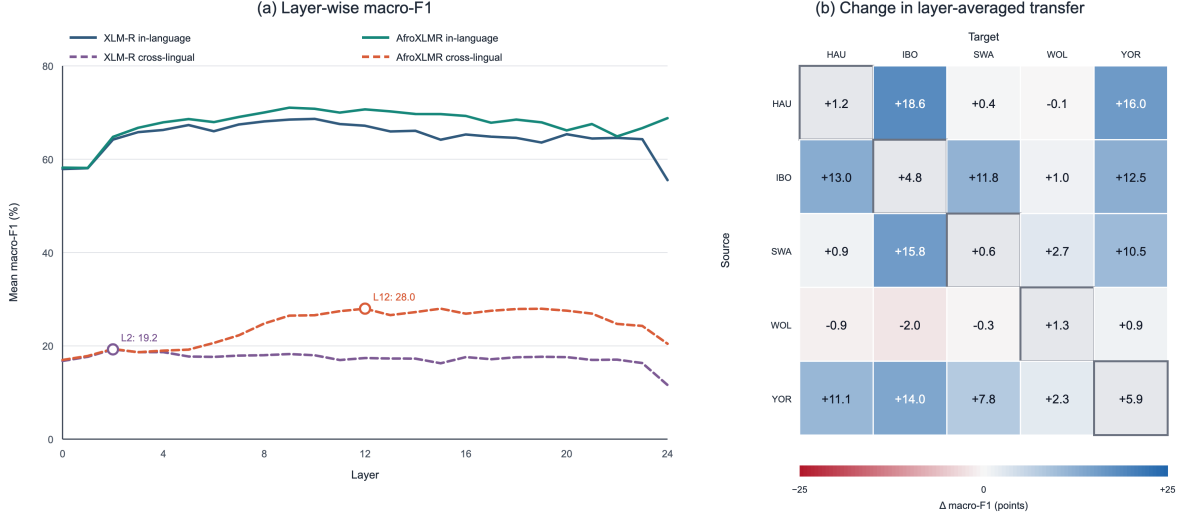


Figure 2: POS accessibility and cross-lingual transfer before and after continued pretraining. Left: mean layer-wise macro-F1 for in-language evaluation (diagonal source–target pairs) and cross-lingual evaluation (off-diagonal pairs). Right: AfroXLMR-minus-XLM-R change in each source–target pair’s macro-F1 averaged across all 25 hidden-state levels; outlined diagonal cells denote in-language evaluation. Adaptation improves cross-lingual transfer across much of the network. Descriptively, mean test performance peaks at layer 2 for XLM-R-large and layer 12 for AfroXLMR-large.

### 5.3 RQ3: How Does Continued Pretraining Reshape Representation Geometry?

Figure 3 compares the layer-wise geometry of the pretrained and adapted models. Continued pretraining generally increases effective dimensionality, but the expansion is concentrated at architecture-specific depths rather than occurring uniformly throughout the network.

AfroXLMR increases effective rank in 88% of language–layer combinations, with the largest average gain of 18.88 dimensions at layer 15. The strongest expansion occurs across the middle encoder layers. AfriqueGemma shows a smaller and more localized effect, with positive changes in 75% of language–layer combinations and a maximum gain of 14.57 dimensions at layer 12. AfriqueLlama exhibits the most widespread expansion: effective rank increases in 94% of language–layer combinations and peaks at layer 25 with a gain of 19.81 dimensions.

The effective-rank increases are accompanied by changes in the distribution of representational variance. In AfroXLMR, for example, the number of principal components required to explain 90% of the variance increases from 118.4 to 143.4 at layer 14, while the variance explained by the leading component falls from 0.138 to 0.068. Mean PC90 increases by 7.70 components for AfriqueGemma and 24.14 for AfriqueLlama;

AfriqeLlama also shows a 0.068 reduction in top-component variance. These results indicate that adaptation generally distributes variance across a broader set of latent directions.

Geometric change is substantially larger for African languages than for the English and French controls. Mean effective-rank gain is 12.03 for African languages compared with 0.66 for the controls in XLM-R, and 17.97 compared with 0.91 in Llama. Gemma shows a smaller difference of 5.42 versus 3.05. Local structure also changes: nearest-neighbour overlap is high at the embedding layer (0.845–0.943) but decreases considerably within the transformer, reaching 0.225 in the final XLM-R layer, 0.342 around Gemma layer 11, and 0.416 in the final Llama layer. Continued pretraining therefore alters both the dimensional structure of the representation space and the local organization of individual examples.

## 6 Discussion

**Continued pretraining selectively reorganizes multilingual representations.** Across all three model families, continued African-language pretraining leaves the embedding representations largely stable while producing substantially greater changes within the contextual layers. However, the location of these changes differs across architectures: XLM-R changes most

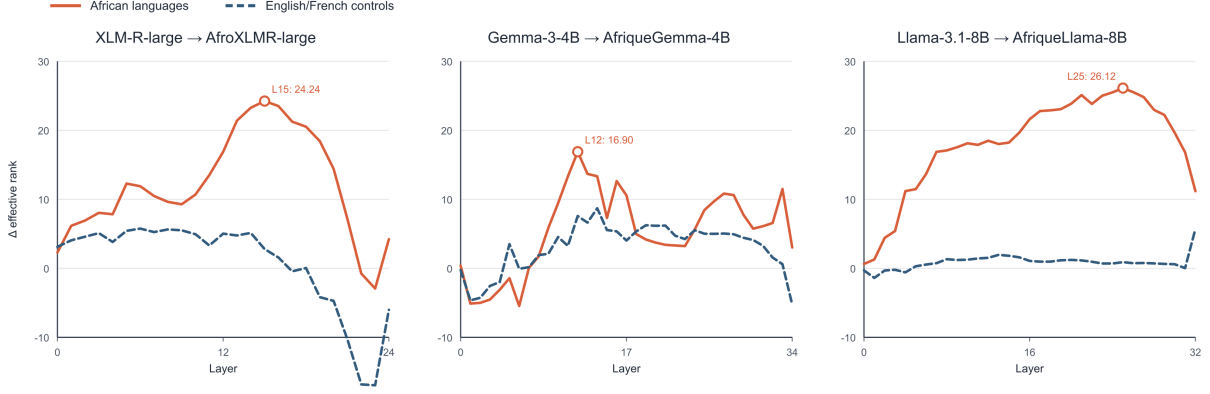


Figure 3: Adapted-minus-base change in effective rank across layers, averaged separately over six African languages and the English/French controls. Positive values indicate that variance is distributed across more latent directions after continued pretraining. Expansion is strongest in the middle layers of XLM-R, layer 12 of Gemma, and later layers of Llama, and is consistently larger for the African-language group.

strongly in its upper layers, Gemma in its early-to-middle layers, and Llama progressively toward its later layers. The stronger changes observed for African languages relative to the English and French controls further suggest that adaptation is concentrated on the linguistic domain emphasized during continued pretraining. Together, these results indicate that there is no single transformer depth at which multilingual adaptation takes place.

**Representational drift and geometric expansion capture different changes.** Representation similarity and geometry provide complementary views of this reorganization. In AfroXLMR, for example, CKA drift peaks at layer 23, whereas effective-rank expansion peaks at layer 15; some upper layers also exhibit substantial drift despite decreasing effective rank. Representational change therefore does not necessarily imply dimensional expansion. CKA characterizes how far an adapted representation departs from its pre-trained counterpart, while effective rank and spectral measures describe how the resulting representation space is internally organized.

**Continued pretraining reorganizes transferable linguistic knowledge.** The probing results provide a functional counterpart to these structural changes. Cross-lingual POS transfer improves substantially more than in-language probing, while the layer of maximum transfer shifts from the lower layers of XLM-R to the intermediate layers of AfroXLMR. This suggests that continued pretraining changes not only the accessibil-

ity of morphosyntactic information, but also where transferable linguistic information is represented across the network. The gains are nevertheless language dependent, with particularly large improvements for transfer into Igbo and Yoruba and considerably smaller changes for Wolof; our experiments do not determine whether these differences arise from pretraining data coverage, tokenization, linguistic relatedness, or properties of the evaluation data.

**Implications for multilingual adaptation.** These findings suggest that continued pretraining should be viewed as a process of selective representational reorganization rather than uniform refinement of an existing multilingual model. The architecture-specific localization of representational change raises the possibility that future adaptation methods could allocate capacity selectively across layers, for example through layer-specific parameter-efficient adaptation or selective freezing. More broadly, the differences between representation similarity, geometry, and cross-lingual transfer show that no single measurement provides a complete account of multilingual adaptation; understanding how a model adapts requires examining both structural changes in its representation space and the functional knowledge those representations support.

## 7 Limitations

English and French provide language controls, but we do not include continued pretraining on a matched non-African corpus. We therefore cannot isolate effects specific to African-language adap-

tation from effects of continued pretraining more generally. Model families also differ in training objective and adaptation recipe, preventing attribution of their differences to architecture alone. Our functional evaluation is limited to POS probing on the encoder pair and does not establish improvements across downstream tasks.

Our CKA analysis compares corresponding layers only and therefore cannot establish layer migration, which requires all-pairs cross-layer CKA. The study covers three model families, eight FLEURS languages, and POS probing for five African languages, so its findings may not generalize beyond these architectures, languages, or tasks. Finally, the observed relationships among representational drift, geometry, and cross-lingual transfer are correlational rather than causal.

## 8 Conclusion

Continued African-language pretraining selectively reorganizes contextual representations and their geometry at different depths across the model pairs studied. For the encoder pair, adaptation also strengthens cross-lingual POS transfer.

## References

- Ife Adebbara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. [SERENGETI: Massively multilingual language models for Africa](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.
- Ife Adebbara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A. Elmadany, and Muhammad Abdul-Mageed. 2025. [Where are we? evaluating LLM performance on African languages](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32704–32731, Vienna, Austria. Association for Computational Linguistics.
- Wisdom Aduah and Francois Meyer. 2025. [Designing and contextualising probes for African languages](#). In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 39–51, Vienna, Austria. Association for Computational Linguistics.
- Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. [Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.
- 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.
- Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. [Pretrained language model embryology: The birth of ALBERT](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, 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. 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, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. [Fleurs: Few-shot learning evaluation of universal representations of speech](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- 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.
- 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.
- Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Ife Adebbara, Jesujoba O. Alabi, Gilles Hacheme, Shamsuddeen Hassan Muhammad, Joyce Nabende, and 1 others. 2023. [Masakhapos: Part-of-speech tagging for typologically diverse african languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10839–10849. Association for Computational Linguistics.

- Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. [AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages](#). In *Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing (SustainLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). 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 55–65, Hong Kong, China. Association for Computational Linguistics.
- Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. 2021. [Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids’ representations](#). In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 375–388, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Gemma Team. 2025. [Gemma 3 technical report](#). Preprint, arXiv:2503.19786.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743. Association for Computational Linguistics.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. [What does BERT learn about the structure of language?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. [Similarity of neural network representations revisited](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.
- Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. [What happens to BERT embeddings during fine-tuning?](#) In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.
- Ari S. Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31.
- Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. [Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages](#). In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. 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.
- Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30.
- Anton Razzhigaev, Matvey Mikhalechuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. [The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 868–874, St. Julian’s, Malta. Association for Computational Linguistics.
- Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. *Euro-pean Signal Processing Conference (EUSIPCO)*.
- Neha Verma, Kenton Murray, and Kevin Duh. 2024. [Exploring geometric representational disparities between multilingual and bilingual translation models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6909–6921, Torino, Italia. ELRA and ICCL.
- 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.

Hao Yu, Tianyi Xu, Michael A. Hedderich, Wassim Hamidouche, Syed Waqas Zamir, and David Ifeoluwa Adelani. 2026. [AfriqueLLM: How data mixing and model architecture impact continued pre-training for African languages](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5909–5928, San Diego, California, United States. Association for Computational Linguistics.

## **A Language-Level Representation Plasticity**

Figure 4 expands the grouped curves in Figure 1 into language-specific profiles. It shows that architecture-level trends coexist with substantial variation across individual languages.

## **B Detailed POS Transfer Results**

For completeness, we also examine the maximum test performance observed across the 25 hidden-state levels. Under this post hoc best-layer analysis, mean cross-lingual macro-F1 increases from 20.8% to 30.1%, a gain of 9.3 points. Because the evaluation set is used to identify the best-performing layer, these values should be interpreted as descriptive upper bounds rather than selection-free estimates.

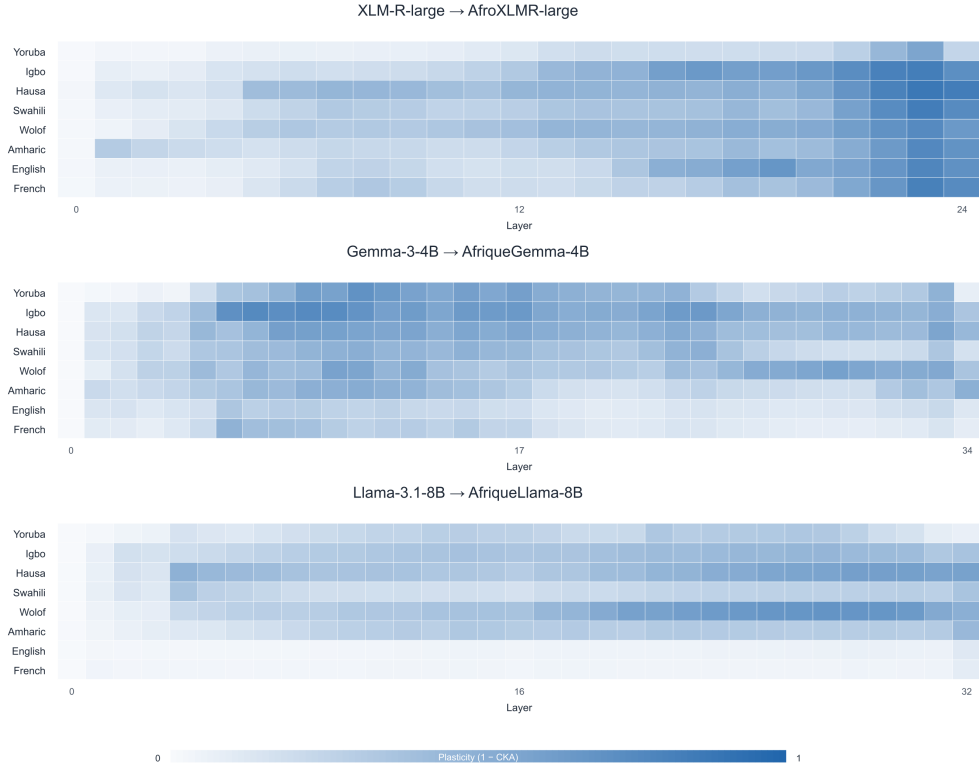


Figure 4: Language-by-layer representation plasticity for all pretrained–adapted model pairs. Each cell reports sentence-level  $1 - \text{CKA}$  for one language and corresponding layer. Darker cells indicate greater representational change. English and French serve as controls.

| Model          | In-language Macro-F1 | Cross-lingual Macro-F1 | In-language Accuracy | Cross-lingual Accuracy |
|----------------|----------------------|------------------------|----------------------|------------------------|
| XLM-R-large    | 64.9                 | 17.3                   | 78.8                 | 35.0                   |
| AfroXLMR-large | 67.6                 | 24.1                   | 81.9                 | 43.2                   |
| Adapted – base | +2.8                 | +6.8                   | +3.1                 | +8.2                   |

Table 2: Aggregate layer-averaged POS probe performance (%). Performance is first averaged across all 25 hidden-state levels for each source–target pair and then averaged across diagonal pairs for in-language evaluation and off-diagonal pairs for cross-lingual evaluation.



Figure 5: Post hoc best-layer POS macro-F1 for every source–target direction in XLM-R-large and AfroXLMR-large. Rows identify the probe-training language and columns identify the evaluation language; outlined diagonal cells are in-language evaluations. Both panels use the same color scale. Because the test set is used to identify each pair’s best-performing layer, these values are descriptive upper bounds rather than selection-free estimates.