

Literature Review: Post-Training Learning Dynamics of Pretrained Language Models for South African Languages

Siyabonga Madondo
University of Cape Town
Cape Town, South Africa
MDNSIY014@myuct.ac.za

ABSTRACT

Pretrained language models (PLMs) are widely adapted to new languages and tasks through continued pretraining and fine-tuning, collectively known as post-training. Despite their use in low-resource contexts, the learning process underlying this adaptation remains poorly understood. Learning dynamics provides a way to study this process by analysing model behaviour at different stages of training to understand how linguistic knowledge develops. However, research on post-training learning dynamics is limited, particularly for low-resource South African languages. This review surveys relevant literature, including pretrained transformer architectures, transfer learning and language adaptation, NLP for South African languages, and methods for analysing learning dynamics. We conclude by identifying key gaps that motivate future research.

KEYWORDS

Natural Language Processing, Learning Dynamics, Transfer Learning, Large Language Models, Continued Pretraining, AI Interpretability

1 INTRODUCTION

Pretrained Language Models (PLMs), trained on large amounts of text, have become the foundation of modern NLP, showing strong performance in many generation and understanding tasks [12, 24, 35]. These PLMs can be adapted to new languages and tasks through continued pretraining and fine-tuning, which are standard approaches for model adaptation [18, 19]. Together, these adaptation steps are referred to as post-training. Modern PLMs are based on deep neural networks, specifically multi-layer transformer architectures [41], which contain a large number of parameters that interact non-linearly across multiple layers. Their complex structure makes it difficult to understand how individual predictions are produced [5]. It also remains largely unclear how linguistic knowledge is distributed across the network and when it develops during training [28]. Understanding model adaptation is especially important for low-resource languages, where training data is limited and effective adaptation strategies are crucial. PLMs generally perform worse on low-resource languages than on high-resource languages [1], but due to their limited interpretability, it is not always clear why or where they fail. Improving our understanding of how linguistic knowledge develops during adaptation could inform the design of more effective models and strategies for low-resource languages.

Many African languages, including South African languages such as isiXhosa, isiZulu, and Sesotho, have small digital corpora and limited NLP resources compared to high-resource languages, such as English [1, 27]. Although recent work has expanded datasets and

resources for some South African languages, the overall availability of digital text in these languages remains limited. For example, the Nguni language group represents a large proportion of the South African population, yet publicly available datasets for these languages are scarce [27]. Multilingual PLMs, trained on multilingual corpora which cover many languages, attempt to address this through cross-lingual transfer, where knowledge from high-resource languages can improve performance in low-resource environments [11]. However, the effectiveness of this transfer depends on factors such as language similarity and the target language's representation in the pretraining data [21, 22]. For languages that are poorly represented, cross-lingual transfer alone is often insufficient. One of the most established ways to improve PLM performance for low-resource languages is continued pretraining, which further trains a multilingual model on target language data [18]. This approach leverages the cross-lingual knowledge already acquired during multilingual training, while further specialising the model for the target language [3]. However, effective adaptation remains difficult without understanding how linguistic knowledge develops during this process.

One way to better understand how models learn during adaptation is to study their learning dynamics. Rather than evaluating a model only after training, researchers examine its behaviour at different checkpoints [15, 28, 36]. Observing models in this way allows researchers to see which types of knowledge are learned first, which appear later, and how quickly they are acquired [9, 23]. Beyond improving our understanding of model behaviour, such analyses can identify when and why learning breaks down, and inform more efficient training by revealing which types of knowledge require more exposure and which are acquired quickly [23]. Notably, if some linguistic knowledge is learned early in training, models may require fewer iterations to reach sufficient levels of competence in those areas [23], reducing training time and computational cost. This is particularly relevant in post-training, where understanding how models learn during adaptation may inform more effective approaches for new languages. While recent work has examined the internal representations of fully adapted models in South African languages to analyse what linguistic information they encode [2], how this knowledge develops during the adaptation process remains largely unexplored.

This review surveys literature related to post-training learning dynamics for low-resource South African languages, covering pretrained transformer architectures, transfer learning and language adaptation, NLP for South African languages, and methods for studying learning dynamics. We conclude by highlighting gaps in prior work that motivate this research direction.

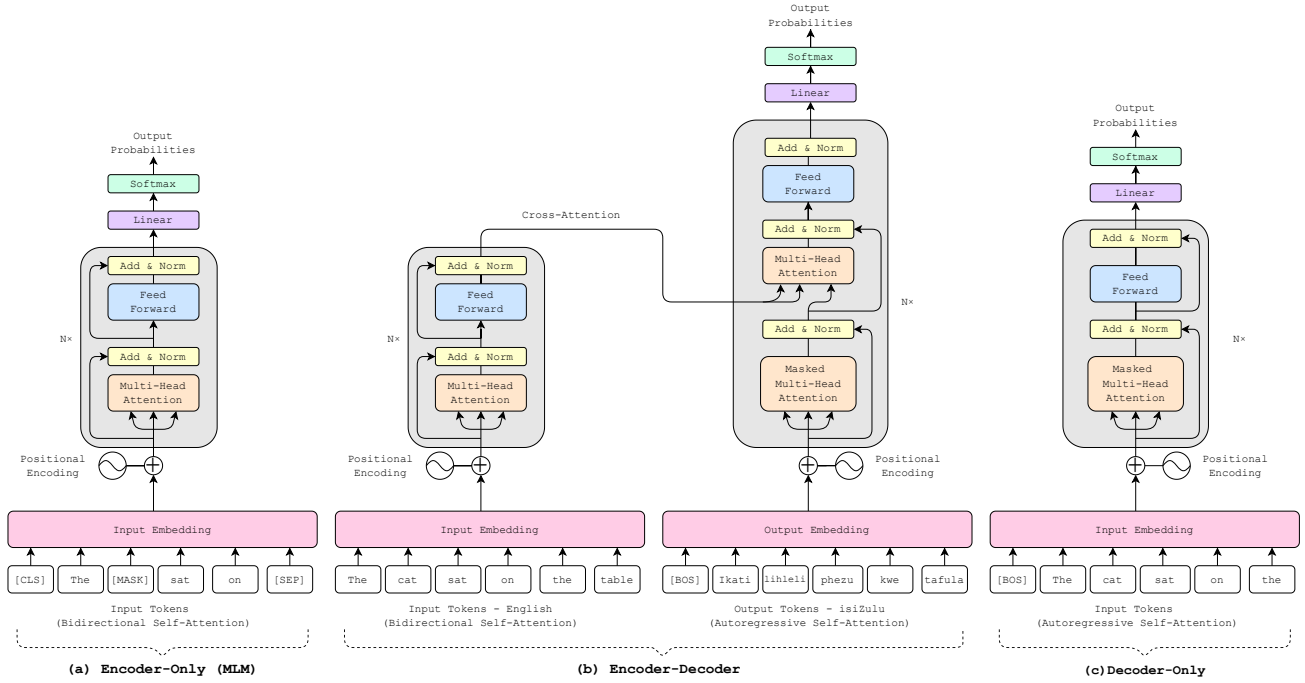


Figure 1: Illustration of the three Transformer architectures, adapted from Vaswani et al. [2017]. (a) Encoder-only models use bidirectional self-attention over the full input sequence. (b) Encoder-decoder models combine both components, with the decoder attending to encoder outputs via cross-attention. (c) Decoder-only models use masked causal self-attention for autoregressive generation. Note that exact architectural details vary across specific models such as BERT, T5 and GPT.

2 PRETRAINED TRANSFORMERS

The architectural design of PLMs directly influences how they acquire and represent linguistic knowledge, making it a necessary foundation for studying their post-training learning dynamics. Modern PLMs are built on the transformer architecture [41], which uses self-attention to capture contextual relationships between all tokens in a sequence. Based on their architecture, they can be broadly categorised into masked language models (MLMs), encoder-decoder models, and decoder-only models, each suited to different NLP tasks. This section introduces each model class, highlighting key examples and their relevance for studying learning dynamics during post-training, as summarised in Table 1 and illustrated in Figure 1.

2.1 Masked Language Models

MLMs are pretrained transformers primarily used for natural language understanding (NLU) tasks, which require comprehending the meaning of input text [13]. Unlike autoregressive models, which predict the next token in a sequence, MLMs use a masked language modelling objective. In this setup, a proportion of the input tokens is randomly masked, and the model is trained to predict them based on the surrounding context [12], as shown in Figure 1. This bidirectional training objective allows MLMs to develop rich contextual representations of language, making them well-suited for tasks such as named entity recognition (NER), part-of-speech (POS) tagging, and text classification [12, 27]. These representations support analyses of how linguistic knowledge develops over time.

2.1.1 BERT. BERT, introduced by Devlin et al. [2018], is a unified architecture that popularised the MLM pretraining objective for bidirectional contextual representation learning. Unlike previous autoregressive models, BERT considers context from both directions simultaneously, enabling richer contextual representations. During pretraining, 15% of input tokens are masked. Of these, 80% are replaced with a special [MASK] token, 10% with a random token, and 10% are left unchanged. This strategy mitigates the mismatch between pretraining and fine-tuning, since the [MASK] token does not appear during fine-tuning, while also forcing the model to maintain a consistent representation of every token. BERT is also pretrained on a next sentence prediction (NSP) objective, where the model learns to predict whether two sentences naturally follow each other. After pretraining, its unified architecture allows BERT to be fine-tuned on downstream NLU tasks with minimal architectural changes [12]. Devlin et al. [2018] also introduced mBERT, a multilingual variant trained on text from 104 languages using the same architecture.

2.1.2 RoBERTa. RoBERTa, proposed by Liu et al. [2019], revisits the BERT pretraining procedure and identifies several areas for improvement. RoBERTa removes the NSP objective, finding that it does not consistently improve BERT’s downstream performance. It also adopts dynamic masking, generating a new masking pattern for each sequence fed to the model. This contrasts with BERT’s static masking strategy, which masks the same tokens throughout training. By training on larger datasets with larger batch sizes for

more steps, Liu et al. [2019] demonstrate that BERT was significantly undertrained, and that these modifications lead to consistent performance improvements across several NLU benchmarks.

2.1.3 XLM-R. XLM-R, presented by Conneau et al. [2020], adapts the RoBERTa architecture to the multilingual setting. It was pre-trained on 2.5 terabytes of filtered CommonCrawl data across 100 languages, including 1.3 GiB of Afrikaans and 0.1 GiB of isiXhosa, using an unsupervised MLM objective with no paired or translated data during training. To handle language diversity, XLM-R uses a large shared vocabulary with SentencePiece tokenisation and omits language embeddings, allowing the model to generalise across languages more effectively. XLM-R significantly outperforms its predecessor mBERT across a range of cross-lingual benchmarks [11]. Both models learn representations where semantically similar concepts from different languages are mapped to similar positions in the embedding space [32]. Languages that share typological or structural similarities can reinforce each other during multilingual pretraining [32], which is relevant for related languages such as the Nguni languages. This enables zero-shot cross-lingual transfer, where a model fine-tuned on a task in one language can perform the same task in other languages without additional training [11]. XLM-R is relevant to this review, as it serves as the base model for several PLMs adapted for South African languages, as discussed in Section 3.

2.2 Encoder-Decoder Models

While MLMs are primarily designed for understanding text, encoder-decoder models focus on natural language generation (NLG), where text is generated from an input sequence [35]. They are trained using a sequence-to-sequence (Seq2Seq) objective, where a bidirectional encoder processes the full input, and an autoregressive decoder generates the output token-by-token based on the encoder’s representations [41], as shown in Figure 1. This structure supports tasks such as machine translation (MT), abstractive summarisation (SUM), question answering (QA), and data-to-text (D2T) generation [27, 35]. This makes them useful for studying how generative capabilities develop during post-training, complementing MLM-based analyses of language understanding.

2.2.1 T5. T5, proposed by Raffel et al. [2023], introduced a unified text-to-text framework where every NLP task is framed as converting an input text string to an output text string. This allows a single model to be pretrained and fine-tuned consistently across diverse tasks. T5 is pretrained using a span corruption objective, where 15% of input tokens are randomly dropped, and consecutive spans of dropped tokens are replaced with unique placeholder tokens, called sentinel tokens. These tokens mark removed text. The model is then trained to reconstruct the original dropped spans, with the target sequence consisting of the dropped spans delimited by the same sentinel tokens. Raffel et al. [2023] evaluate T5 across 24 tasks, demonstrating strong performance across language understanding benchmarks. T5 serves as the foundation for several multilingual and byte-level extensions relevant to this review, discussed below.

2.2.2 mT5. mT5, introduced by Xue et al. [2021], is a multilingual extension of T5 pretrained on mC4, a multilingual dataset derived from Common Crawl covering 101 languages, including isiZulu,

isiXhosa, and Afrikaans. It adopts the same text-to-text framework, encoder-decoder architecture, and span corruption objective as T5. Unlike XLM-R, mT5 retains full generative capabilities, enabling cross-lingual text generation across tasks such as translation, summarisation, and question answering. However, multilingual generative models can suffer from accidental translation in zero-shot contexts, where predictions are partially generated in the wrong language. Xue et al. [2021] address this by mixing a small amount of unlabelled pretraining data into the fine-tuning stage. mT5 has been adapted for several African languages, which is further discussed in Section 3.

2.2.3 ByT5. Building on mT5, ByT5, proposed by Xue et al. [2022], operates directly on raw bytes rather than subword tokens, removing the need for a predefined subword tokeniser. This makes it robust to noise and avoids the problem of subword tokenisers fragmenting morphologically complex words in ways that hide their grammatical structure. To handle longer byte sequences, ByT5 uses a deeper encoder than its decoder. These properties make ByT5 well-suited for agglutinative and low-resource languages, where standard tokenisation often struggles to capture morphological structure [26, 31]. It has been adapted for several South African languages, as discussed in Section 3.

2.3 Decoder-Only Models

The final architectural class we discuss is decoder-only models, used primarily for text generation tasks. They consist of a single autoregressive decoder that generates text token by token, conditioning each prediction on all preceding tokens through masked self-attention [33, 41], as shown in Figure 1. While Radford et al. [2018] initially demonstrated that generative pretraining transfers effectively to NLU tasks, such as question answering and sentiment analysis, subsequent work has shown its suitability for open-ended generation, dialogue, and code generation [7, 34]. Large-scale training and in-context learning (ICL) have made decoder-only models the dominant architecture in modern NLP, as seen in the GPT series [7, 33, 34]. They also provide a useful framework for studying how generative capabilities develop during post-training.

2.3.1 GPT. The GPT model, introduced by Radford et al. [2018], established the decoder-only pretraining paradigm for NLG tasks. GPT is pretrained using a causal language modelling (CLM) objective, where the model learns to predict the next token based on all preceding tokens. Later iterations, notably GPT-2 [34] and GPT-3 [7], demonstrated that scaling decoder-only models to billions of parameters improves few-shot and zero-shot performance across a range of NLP tasks, including question answering and arithmetic, without task-specific fine-tuning.

The models discussed in this section form the architectural foundation for the language-adapted models reviewed in Section 3. The choice of architecture determines the pretraining objective, the types of linguistic knowledge a model can acquire, and the downstream tasks it can be adapted to, all of which influence how learning dynamics are studied during post-training.

Table 1: Summary of PLMs, their architectures, pretraining objectives, tokenisation, and tasks. ✓ indicates models explicitly trained for multilingual use; ✓* indicates models with multilingual capacity but primarily trained on English data.

Model	Type	Objective	Tokenisation	Tasks	Multilingual
BERT	MLM	MLM + NSP	WordPiece	NER, POS, Classification	
RoBERTa	MLM	MLM	BPE	NER, POS, Classification	
XLM-R	MLM	MLM	SentencePiece	NER, POS, Classification	✓
T5	Enc-Dec	Span Corruption	SentencePiece	MT, QA, SUM	
ByT5	Enc-Dec	Span Corruption	Bytes	MT, D2T	✓*
mT5	Enc-Dec	Span Corruption	SentencePiece	MT, QA, SUM	✓
GPT	Decoder	CLM	BPE	QA, Generation	

3 TRANSFER LEARNING IN NLP

Transfer learning has become a standard approach for adapting PLMs to new languages and tasks [18, 19, 35]. Instead of pretraining a model from scratch, it uses knowledge acquired during pretraining and applies it to new contexts [19]. In NLP, transfer learning typically takes two forms. The first is language adaptation, where a model is extended through continued pretraining on new language data [1, 18]. The second is task transfer, where a pretrained model is fine-tuned on labelled data for a specific downstream task [19]. Both approaches are important for low-resource South African languages, where limited data makes pretraining models challenging [3, 27]. This section examines language adaptation and task-specific fine-tuning strategies, with a focus on how these approaches are applied to South African languages.

3.1 Language Adaptation

Language adaptation extends a pretrained model to new language data through continued pretraining, where an existing model is further trained on target language text to develop stronger representations of that language [18]. This is a standard approach for adapting multilingual PLMs to low-resource languages, where insufficient data often makes training from scratch impractical [1, 11].

Several unannotated corpora exist for continued pretraining on South African languages. The mC4 corpus [45] is a large multilingual web crawl covering many languages, though its scale comes at the cost of lower text quality [27]. The NCHLT Text Corpora [14] provide higher-quality text drawn from South African government documents and domain-specific prose. WURA [30] is a curated multilingual corpus for African languages, created by filtering mC4 to remove low-quality text. Recently, curated corpora such as PuoData [25] have been developed specifically for low-resource South African languages, prioritising quality over scale. However, coverage remains uneven across languages and is low relative to high-resource languages, which remains an adaptation challenge.

An alternative to language-specific adaptation is multilingual pretraining, where a model is trained on corpora covering many languages simultaneously [11]. This approach is discussed in Section 2 with XLM-R [11] and mT5 [45], both trained on over 100 languages. Such broad coverage improves performance for low-resource languages through cross-lingual transfer from high-resource languages, but it can reduce performance for individual languages. This trade-off is known as the curse of multilinguality [11]. Consequently,

language-specific models obtained through continued pretraining, such as Afro-XLMR [3] and Nguni-XLMR [27], tend to outperform general multilingual models on their target languages. Understanding these trade-offs is important for designing effective adaptation strategies for low-resource languages.

3.2 Task-Specific Adaptation

Once a model has been pretrained and optionally adapted to a target language, it can be fine-tuned on labelled data for a specific downstream task [18, 19]. Its parameters are updated using task-specific examples, allowing the model to specialise for tasks such as NER, POS tagging, and text generation [1, 3, 27]. Prior work has shown that combining continued pretraining and fine-tuning leads to consistent improvements in downstream performance, including in low-resource contexts [1, 18].

Model performance on downstream tasks is typically evaluated using benchmarks, which are collections of tasks and datasets designed to measure performance [42]. Several benchmarks have been developed to evaluate model performance on African languages. AfroBench [29] provides a multilingual evaluation framework covering 15 NLP tasks across 22 datasets and 64 African languages. AfriMTEB [40] is a multilingual text embedding benchmark adapted from MMTEB [16], designed to address the under-representation of African languages by evaluating semantic representations across 59 languages, 14 tasks, and 38 datasets. For South African languages, Meyer et al. [2024] introduced NGLUEni, the first centralised evaluation suite for the Nguni languages, covering 6 NLU and NLG tasks across isiXhosa, isiZulu, isiNdebele, and Siswati. These benchmarks are important for analysing learning dynamics, as they enable researchers to evaluate model performance over time and gain insight into how models learn during adaptation.

These strategies and evaluation benchmarks provide the foundation for analysing post-training learning dynamics, as they define both how models are adapted and how their performance is measured. Table 2 summarises PLMs adapted for African languages using continued pretraining, along with their coverage across South African languages.

4 LEARNING DYNAMICS

Learning dynamics is the study of how models acquire knowledge throughout training. Instead of evaluating a model only after training has completed, learning dynamics research analyses model

Table 2: South African and African language PLMs adapted using continued pretraining and the languages they cover. Languages include Afrikaans (af), isiNdebele (nr), Sesotho (st), Siswati (ss), Setswana (tn), isiXhosa (xh), isiZulu (zu). Note that isiXhosa and isiZulu are the most consistently represented languages across models, while several languages, including isiNdebele and Siswati, have limited coverage and others, such as Tshivenda and Sepedi, have no dedicated PLMs. This uneven coverage reflects broader data and resource constraints in South African NLP.

Model	af	nr	ss	st	tn	xh	zu	Authors
Masked Language Models (MLMs)								
Afro-XLMR	✓			✓		✓	✓	Alabi et al. [2022]
Nguni-XLMR		✓	✓			✓	✓	Meyer et al. [2024]
Encoder-Decoder Models								
Afri-ByT5	✓			✓		✓	✓	Adelani et al. [2022]
Afri-mT5	✓			✓		✓	✓	Adelani et al. [2022]
Nguni-ByT5		✓	✓			✓	✓	Meyer et al. [2024]
Decoder-Only Models								
AfriqueLLM	✓			✓	✓	✓	✓	Yu et al. [2026]
Lugha-Llama	✓					✓	✓	Buzaaba et al. [2025]

behaviour at multiple checkpoints to observe which capabilities emerge first, which develop later, and how quickly different types of knowledge are acquired [28, 36]. This approach has been applied successfully to analyse how models develop syntactic, semantic, reasoning, and world knowledge during pretraining [9, 23, 28]. In contrast, research on learning dynamics during post-training remains limited. While recent work has examined the internal representations of fully adapted models to see what linguistic information they encode [2], the development of this knowledge during the adaptation process remains largely unexplored, with only very recent work beginning to address this [15]. To provide context for our work, this section surveys the methods used to study learning dynamics and the prior work that has applied them.

4.1 Methods for Studying Learning Dynamics

Learning dynamics research uses various methods to study how models acquire knowledge during training, each capturing different aspects of the learning process over time. Broadly, these methods can be grouped into fine-tuning on downstream tasks, probing internal representations, zero-shot evaluations, and measuring internal representations. This section introduces each approach and reviews how it has been applied in prior work.

4.1.1 Fine-Tuning on Downstream Tasks. Fine-tuning on downstream tasks involves taking intermediate model checkpoints and training them on labelled data for specific tasks [9, 23]. Unlike loss-based performance measures, which can remain stable even as underlying representations continue to shift [28], or appear to improve while catastrophic forgetting occurs simultaneously [15], this method reveals how well the model can specialise for specific tasks at each stage of training. By tracking task performance as a function of training steps, researchers can identify the point at which representations are sufficient for effective task specialisation [9, 23]. This approach is relevant to this work, as it provides

a direct measure of how adaptation affects task performance over time. The following studies show how it has been applied.

Chiang et al. [2020] investigated the learning dynamics of ALBERT [20] by fine-tuning intermediate checkpoints and found that models are prepared for fine-tuning very early, reaching strong performance on the GLUE [42] benchmark within the first 25% of training updates. These findings suggest that training until pretraining loss converges is often unnecessary, with later performance gains being disproportionate to the additional computational cost. Liu et al. [2021] similarly fine-tuned RoBERTa checkpoints on several benchmarks and showed that simple linguistic tasks reach strong performance within 16–32% of pretraining, while more complex factual, commonsense, and reasoning tasks benefit from significantly longer training. This indicates that different types of knowledge are acquired at different stages, with syntactic knowledge being acquired earlier than world knowledge and reasoning. Consistent with Chiang et al. [2020], they also find that longer pretraining does not always improve fine-tuning performance. Müller-Eberstein et al. [2023] also evaluated fine-tuned performance across checkpoints from MultiBERTs [37], finding that most NLP tasks reach 90% of their final performance when fine-tuned from a checkpoint that has only seen 0.5% of the total pretraining data. This again supports the finding that task competence is acquired early in pretraining, raising the question of whether similar patterns hold during continued pretraining on low-resource languages.

4.1.2 Probing. Probing involves training a lightweight classifier on top of fixed intermediate representations to measure the extent to which specific linguistic features are encoded [23]. It isolates information already present in the hidden layers without updating model weights, revealing what a model knows and when that knowledge is acquired during training, if measured over time. Following Belinkov et al. [2020], probes can be categorised into structural and behavioural (zero-shot) probes. Structural probes train a lightweight classifier on top of a model’s internal representations to predict linguistic labels, such as part-of-speech tags. Behavioural

probes, in contrast, directly evaluate the model’s ability to make predictions without introducing additional parameters. The following studies use structural probing to analyse knowledge acquisition over time.

Chiang et al. [2020] applied structural probes using the edge probing framework [38] to evaluate ALBERT’s hidden layer representations throughout pretraining. They found that syntactic and semantic knowledge develop simultaneously in the early stages of pretraining, and that layers near the output acquire knowledge faster than layers closer to the input initially, before eventually surpassing them and decaying towards the end of training. This pattern supports the view that pretraining beyond a certain point may be inefficient. Liu et al. [2021] similarly applied structural probes to RoBERTa checkpoints, with findings consistent with those discussed in Section 4.1.1. Müller-Eberstein et al. [2023] used information-theoretic probes on MultiBERTs to study how representational subspaces change during pretraining. They found that these subspaces keep shifting and specialising until the end of pretraining, even when performance metrics appear to have stabilised. This suggests that loss and accuracy alone do not fully capture changes in the model’s representations. They also found that the traditional NLP pipeline, where syntactic information is encoded in lower layers and semantic information in higher ones, only appears late in pretraining, rather than being present from the start. Whether these patterns hold during continued pretraining for low-resource languages remains an open question.

4.1.3 Zero-Shot Evaluation. Zero-shot evaluations assess a model’s capabilities at different stages of training without task-specific fine-tuning [34]. In learning dynamics research, it is used to track how linguistic, factual, and reasoning abilities develop across checkpoints [15, 17, 23]. Unlike fine-tuning and structural probing, it evaluates the model directly, revealing capabilities learned through the pretraining objective alone. The following studies use zero-shot evaluations to study model capabilities over time.

Choshen et al. [2022] presented models with pairs of sentences consisting of a grammatically correct sentence and an ungrammatical version, and evaluated their ability to assign a higher probability to the correct one. Applying this approach to the BLiMP benchmark [43] across models with different architectures and initialisations, they found that models follow a consistent grammar-learning order regardless of their architecture. Evanson et al. [2023] similarly used zero-shot probes from BLiMP and other benchmarks to compare GPT-2’s acquisition stages to those of human children, finding that both master simple subject-verb agreement before more complex structures. Biderman et al. [2023] evaluated checkpoints from the Pythia suite across various benchmarks and models using zero-shot evaluation, identifying a phase change at roughly 45% of training where larger models begin to show accuracy that is strongly correlated with the frequency of task-relevant terms in the pretraining data. Chiang et al. [2020] and Liu et al. [2021] also used zero-shot evaluations alongside the probing and fine-tuning experiments discussed in Sections 4.1.1 and 4.1.2, finding consistent results across methods.

In contrast to the pretraining-focused studies above, Elhady et al. [2025] investigated learning dynamics during continued pretraining, making their work the most directly relevant to ours. They

adapted Llama 2 7B [39] to Basque, Arabic, and Indonesian by continuing pretraining on target language data, comparing mixtures that included or omitted English. Monitoring zero-shot ICL performance across checkpoints, they identified a critical period in the first few steps of adaptation where catastrophic forgetting occurs if English is omitted, with ICL capabilities dropping to near zero almost immediately. This drop in performance is tied to a rapid shift in model parameters early in training. However, the drop does not manifest in validation perplexity or downstream accuracy until much later. These findings are consistent with earlier observations about critical learning phases during pretraining [9, 28], but in a post-training context. Whether similar dynamics occur for encoder-only and encoder-decoder models adapted to low-resource South African languages remains an open question.

4.1.4 Internal Representation Measures. Internal measures analyse a model’s parameters or hidden representations directly, without requiring labelled data or additional classifiers [28, 36]. These methods are useful for identifying representational shifts that may not be visible through task performance alone. The following studies show how this technique has been applied in prior research.

Saphra and Lopez [2019] used Singular Vector Canonical Correlation Analysis (SVCCA) to compare a model’s internal representations with those of independently trained linguistic taggers across training. They found that recurrent layers, following a brief initial decline, quickly become more aligned with syntactic structure early in training, while embedding layers diverge as they specialise for the language modelling objective. This aligns with findings from probing and fine-tuning, where syntactic knowledge is also acquired early [9, 23]. Müller-Eberstein et al. [2023] used subspace angles to measure the overlap between task-specific regions in the model’s representation space, finding that syntactic and semantic subspaces are closely aligned during the first 0.5% of training before separating into specialised layers. This provides a geometric explanation for the rapid early task competence and later layer specialisation observed through fine-tuning and probing. In a post-training context, Elhady et al. [2025] measured the average change in model weights from their initial state during continued pretraining, finding that rapid parameter divergence occurs within the first 1,000 steps, aligning with the catastrophic forgetting observed in their zero-shot evaluations. It is still unclear whether models show similar dynamics when adapted to low-resource South African languages during continued pretraining.

Overall, these studies show that models acquire simpler knowledge, such as syntax, early in training, while more complex reasoning and factual knowledge develop gradually. Performance metrics alone do not fully capture these changes, because internal representations can shift or degrade without affecting observable outcomes. Most research has examined models trained from scratch, leaving open the question of how knowledge develops during adaptation to low-resource languages. This gap motivates our work.

5 Critical Comparison of Prior Work

Continued pretraining has proven to be the most consistent approach for adapting PLMs to low-resource African languages [1, 18], generalising well across successive model generations. This trend

holds across model architectures: in MLMs, continued pretraining of XLM-R produced Afro-XLMR and Nguni-XLMR, both achieving strong performance on a range of NLU tasks for African languages. Encoder-decoder models such as Afri-mT5 and more recent decoder-only models such as AfriqueLLM have similarly benefited from continued pretraining. Across these generations, continued pretraining consistently yields meaningful performance improvements for languages such as isiXhosa, isiZulu, and Sesotho, even when pretraining data is scarce. Despite this sustained success, almost no work has examined continued pretraining from an interpretability perspective. We know that it works, but we do not understand why, when it works best, or where and why it fails.

Understanding why continued pretraining works requires looking beyond final performance metrics to how models acquire knowledge during adaptation. Learning dynamics research provides a method for studying this. By analysing models at multiple checkpoints using techniques such as fine-tuning on downstream tasks, probing internal representations, zero-shot evaluation, and measuring changes in model parameters, researchers can observe which types of knowledge are acquired first, how stable they are, and which parts of the model specialise for specific tasks. Studies during pretraining have consistently shown that syntactic knowledge is acquired early, while factual and reasoning capabilities develop later [9, 23, 28]. Importantly, training loss alone is a poor indicator of capability. Models can appear to improve while simultaneously losing previously acquired abilities, as demonstrated by catastrophic forgetting during continued pretraining [15]. Applying this approach to continued pretraining for South African languages could help explain when adaptation is most effective, where it fails, and why. These questions remain largely unexplored.

A further gap in existing learning dynamics research is that almost all studies have examined models trained from scratch. Even for high-resource languages, this is fundamentally different from continued pretraining. When training from scratch, a model must acquire all linguistic knowledge from random initialisation [9]. In continued pretraining, the model already encodes substantial knowledge from its initial pretraining, and adaptation involves building on and redirecting existing representations rather than acquiring them from scratch. The learning dynamics of these two settings are therefore expected to differ substantially. This difference may be most noticeable for morphologically complex languages such as isiXhosa and isiZulu, where agglutinative structure and limited pretraining data mean that adaptation must be efficient. For these languages, understanding how knowledge develops during continued pretraining is both practically important and largely unstudied. Applying learning dynamics methods such as fine-tuning intermediate checkpoints and tracking parameter shifts to continued pretraining for South African languages would provide direct insight into how and when adaptation succeeds or fails, informing more effective training strategies for these languages. This review identifies this gap as a key motivation for future work.

6 CONCLUSION

This review has surveyed literature on post-training learning dynamics for low-resource South African languages. We first introduced the transformer architectures that form the foundation of

modern PLMs, then examined how continued pretraining and fine-tuning have been used to adapt these models to low-resource African languages. Finally, we reviewed the main methods used to study learning dynamics and highlighted prior work that applied these techniques during pretraining.

A consistent finding across this literature is that performance metrics alone are insufficient to capture the full picture of how models learn. Internal representations can shift or degrade without any observable change in loss or accuracy, and critical periods of learning and forgetting can occur early in training before manifesting in downstream performance. Despite the sustained success of continued pretraining across successive model generations, very little work has examined this process from a learning dynamics perspective.

This gap is particularly important for low-resource South African languages. The morphological complexity of languages such as isiXhosa and isiZulu, combined with limited pretraining data, suggests that adaptation dynamics may differ substantially from those observed in high-resource languages. Analysing these dynamics through checkpoint-based methods provides a valuable opportunity to better understand how continued pretraining works, when it is most effective, and how it can be optimised for low-resource contexts.

REFERENCES

- [1] David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, Andre N. Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin A. Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire M. Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmunmin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3053–3070. doi:10.18653/v1/2022.naacl-main.223
- [2] Wisdom Aduah and Francois Meyer. 2025. Designing and Contextualising Probes for African Languages. (5 2025). <http://arxiv.org/abs/2505.10081>
- [3] 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*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4336–4349. <https://aclanthology.org/2022.coling-1.382/>
- [4] Yonatan Belinkov, Sebastian Gehrmann, Google Ai, and Ellie Pavlick. 2020. Interpretability and Analysis in Neural NLP. (7 2020), 1–5. doi:10.18653/v1/2020.acl-tutorials.1
- [5] Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics* 7 (1 2019), 49–72. <http://arxiv.org/abs/1812.08951>
- [6] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shrivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *Proceedings of Machine Learning Research* 202 (5 2023), 2397–2430. <http://arxiv.org/abs/2304.01373>
- [7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

- Askill, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Matusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems 2020-December* (5 2020). <https://arxiv.org/pdf/2005.14165>
- [8] Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025. Lugh-Llama: Adapting Large Language Models for African Languages. (4 2025). <http://arxiv.org/abs/2504.06536>
- [9] Cheng-Han Chiang, Sung-Feng Huang, and Hung yi Lee. 2020. Pretrained Language Model Embryology: The Birth of ALBERT. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (10 2020), 6813–6828. <http://arxiv.org/abs/2010.02480>
- [10] Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The Grammar-Learning Trajectories of Neural Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1 (4 2022), 8281–8297. <http://arxiv.org/abs/2109.06096>
- [11] 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. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (4 2020), 8440–8451. <http://arxiv.org/abs/1911.02116>
- [12] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (10 2018), 4171–4186. <https://arxiv.org/pdf/1810.04805>
- [13] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Short-cut Learning of Large Language Models in Natural Language Understanding. *Commun. ACM* 67 (12 2023), 110–120. Issue 1. doi:10.1145/3596490
- [14] Roald Eiselein and Martin Puttkammer. 2014. Developing Text Resources for Ten South African Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declercq, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Reykjavik, Iceland, 3698–3703. <https://aclanthology.org/L14-1106/>
- [15] Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. Emergent Abilities of Large Language Models under Continued Pretraining for Language Adaptation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1 (9 2025), 32174–32186. <http://arxiv.org/abs/2506.00288>
- [16] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, Genta Indra Winata, Saba Sturua, Rafal Poswiata, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatina, Ömer Çagatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pawha, Rafal Poswiata, Kranthi G.V. Kiran, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Votolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. *13th International Conference on Learning Representations, ICLR 2025* 19 (2 2025), 102004–102060. <https://arxiv.org/pdf/2502.13595>
- [17] Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2023), 12205–12218. doi:10.18653/v1/2023.findings-acl.773
- [18] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (4 2020), 8342–8360. doi:10.18653/v1/2020.acl-main.740
- [19] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (5 2018), 328–339. <http://arxiv.org/abs/1801.06146>
- [20] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *8th International Conference on Learning Representations, ICLR 2020* (2 2020). <http://arxiv.org/abs/1909.11942>
- [21] Anne Lauscher, Vinit Ravishanker, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (5 2020), 4483–4499. <http://arxiv.org/abs/2005.00633>
- [22] 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. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (6 2019), 3125–3135. <http://arxiv.org/abs/1905.12688>
- [23] Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing Across Time: What Does RoBERTa Know and When? *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021* (9 2021), 820–842. <http://arxiv.org/abs/2104.07885>
- [24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (7 2019). <http://arxiv.org/abs/1907.11692>
- [25] Vukosi Marivate, Moseli Mots'Oehli, Valencia Wagner, Richard Lastrucci, and Isheanesu Dzingirai. 2023. PuoBERTa: Training and evaluation of a curated language model for Setswana. (10 2023). <http://arxiv.org/abs/2310.09141>
- [26] Francois Meyer and Jan Buys. 2025. The Learning Dynamics of Subword Segmentation for Morphologically Diverse Languages. (11 2025). <http://arxiv.org/abs/2511.09197>
- [27] Francois Meyer, Haiyue Song, Abhisek Chakraborty, Jan Buys, Raj Dabre, and Hideki Tanaka. 2024. NGLUEni: Benchmarking and Adapting Pretrained Language Models for Nguni Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 12247–12258. <https://aclanthology.org/2024.lrec-main.1071/>
- [28] Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. Subspace Chronicles: How Linguistic Information Emerges, Shifts and Interacts during Language Model Training. *Findings of the Association for Computational Linguistics: EMNLP 2023* (10 2023), 13190–13208. <http://arxiv.org/abs/2310.16484>
- [29] Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? (6 2025), 19048–19095. <http://arxiv.org/abs/2311.07978>
- [30] Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Odunayo Ogundepo, Abraham Toluwalase Owodunni, David Ifeoluwa Adelani, and Jimmy Lin. 2023. Better Quality Pre-training Data and T5 Models for African Languages. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings* (2023), 158–168. doi:10.18653/v1/2023.emnlp-main.11
- [31] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. *Advances in Neural Information Processing Systems* 36 (10 2023). <http://arxiv.org/abs/2305.15425>
- [32] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is Multilingual BERT? *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (6 2019), 4996–5001. <http://arxiv.org/abs/1906.01502>
- [33] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. (2018). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- [34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI* (2019). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (9 2023), 1–67. <http://arxiv.org/abs/1910.10683>
- [36] Naomi Saphra and Adam Lopez. 2019. Understanding Learning Dynamics Of Language Models with SVCCA. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (4 2019), 3257–3267. doi:10.18653/v1/N19-1329
- [37] Thibault Sellam, Steve Wadlow, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2022. The MultiBERTs: BERT Reproductions for Robustness Analysis. *ICLR 2022 - 10th International Conference on Learning*

- Representations* (3 2022). <http://arxiv.org/abs/2106.16163>
- [38] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (2019), 4593–4601. doi:10.18653/v1/p19-1452
 - [39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. (7 2023). <http://arxiv.org/abs/2307.09288>
 - [40] Kosei Uemura, Miaoran Zhang, and David Ifeoluwa Adelani. 2026. AfriMTEB and AfriE5: Benchmarking and Adapting Text Embedding Models for African Languages. (3 2026). <http://arxiv.org/abs/2510.23896>
 - [41] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (6 2017), 1. <https://arxiv.org/pdf/1706.03762>
 - [42] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop* (2 2019), 353–355. <http://arxiv.org/abs/1804.07461>
 - [43] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics* 8 (2020), 377–392. doi:10.1162/tacl_a_00321
 - [44] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics* 10 (3 2022), 291–306. <http://arxiv.org/abs/2105.13626>
 - [45] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference* (3 2021), 483–498. <http://arxiv.org/abs/2010.11934>
 - [46] 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. (1 2026). <http://arxiv.org/abs/2601.06395>