

Learning Dynamics Of LM Post-Training: A Literature Review

Rakeen Ramdeen
RMDRAK003@myuct.ac.za
University Of Cape Town
Cape Town, South Africa

Abstract

Pretrained language models achieve state-of-the-art performance in natural language processing tasks. Their success is reliant on large, high-quality datasets, which makes it infeasible for low-resource African languages. To address this, researchers use continued pretraining (CPT) to adapt existing models to new languages. While monolingual CPT yields high performance, it requires substantial disk space per language. Multilingual CPT yields competitive performance with efficiency. However, prior research evaluates these adaptation methods only at the end of training, treating the acquisition of linguistic knowledge as a black box. We review the field of learning dynamics, which analyses models at intermediate checkpoints to uncover exactly how and when linguistic capabilities emerge. Prior research establishes a hierarchical timeline for pretrained models: syntax is learned rapidly, followed by the slower acquisition of facts, with complex reasoning emerging last. The learning dynamics of CPT remain largely unstudied, therefore it is unknown if CPT follows this same trajectory.

Keywords

multilingual NLP, continued pretraining, learning dynamics, low-resource languages, South African languages

ACM Reference Format:

Rakeen Ramdeen. 2026. Learning Dynamics Of LM Post-Training: A Literature Review. In *Proceedings of Computer Science Honours (Honours 2026)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

1 Introduction

Natural language processing (NLP) has been transformed by the arrival of the transformer [25] architecture and large language models [24]. Encoder-only [10], encoder-decoder [21], and decoder-only [24] models have achieved state-of-the-art performance in natural language understanding and natural language generation tasks. However, their success is dependent on procuring large, high-quality datasets, which are often composed of high-resource languages such as English. As a result, low-resource languages, especially African languages, are underrepresented in NLP. South African languages, although being low-resource, have seen some efforts to promote inclusivity [17].

Researchers address these issues by using language adaptation methods such as continued pretraining (CPT). CPT allows researchers to adapt already existing pretrained language models (PLMs) using a smaller corpus of the target language. Monolingual [1] and multilingual [3] CPT have both been successful in improving downstream task performance, however, current literature treats the adaptation process as a black box. By only measuring downstream

task performance at the end of training, researchers overlook how and when these models develop linguistic capabilities. Critical questions remain unanswered: "At what specific point during CPT does a model develop syntactic knowledge?" and "When does the model develop semantic understanding?"

Learning dynamics attempts to answer these questions. By analysing models at intermediate checkpoints throughout the pretraining process, prior work has found a hierarchical timeline of knowledge acquisition. Models acquire basic syntactic knowledge rapidly, before slowly learning factual and commonsense knowledge, and then much later developing reasoning [14, 18, 23]. However, prior learning dynamics research is almost entirely based on the pretraining process and not CPT. This creates a gap in our understanding of how PLMs adapt and acquire new linguistic knowledge from new languages.

We explore different transformer architectures from three main categories and their most influential models. We will consider BERT [10] and its masked language modelling objective, T5 [21] and its unification of NLG tasks, and a brief overview of decoder models like LLaMA [24]. We consider language adaptation methods, focusing on transfer learning, cross-lingual transfer, language-agnostic models, and then pivoting to CPT and its two variants: monolingual and multilingual CPT. We explore learning dynamics across four distinct methodologies: vector analysis, structural probing, zero-shot, and downstream task finetuning. We end our review on a critical comparison of prior work.

2 Pretrained Transformer Models

The Transformer architecture, introduced by Vaswani et al. [2017], replaced prior recurrent models by employing a self-attention mechanism. This mechanism enables each token to attend to every other token in a sequence, allowing its internal representation to become semantically rich with long-range contextual dependencies. The original Transformer uses a full encoder-decoder architecture where the encoder processes input tokens into a mathematical representation and the decoder consumes them autoregressively to produce new output tokens one at a time [25]. We explore three structural variants of the Transformer.

2.1 Masked Language Models (Encoder-only Models)

Following the introduction of the Transformer, GPT-1 [20] utilized the decoder component for unsupervised pretraining and supervised finetuning. They demonstrated semi-supervised learning allows models to acquire significant world knowledge and improve downstream task performance for natural language understanding. However, Devlin et al. [2019] argued that this unidirectional architecture hinders the learned representations of the model. Because GPT-1 is limited to a left-to-right attention mechanism, each token

can only attend to previous tokens, preventing the model from capturing the full context of a sentence. To overcome the limitations of unidirectional architecture, Devlin et al. [2019] introduced BERT, which employs an encoder-only architecture using bidirectional attention. This allows BERT to learn representations from both left and right context in all layers, a significant advancement over shallow bidirectional concatenations like ELMo [19]. The pretraining objective BERT uses is masked language modeling (MLM). This involves randomly masking 15 percent of input tokens with a special [MASK] token. The model's objective is to predict the original token that corresponds to the [MASK] token using its bidirectional context of the unmasked sequence. Additionally, BERT incorporates next sentence prediction (NSP), a binary classification task where the model determines if two sentences follow each other, improving the model's understanding of sentence relationships. BERT achieved state-of-the-art performance on 11 NLP tasks including GLUE [26] and SQuAD [22] benchmarks and proved that a single model could outperform task-specific architectures.

RoBERTa (Robustly Optimized BERT Pretraining Approach) [15] is a replication study of BERT that focuses on improving the model's training recipe and design choices. Specifically, the authors trained RoBERTa on significantly more data, with larger batch sizes, and over a longer training period, revealing that the original BERT model was severely undertrained. Next, they replaced BERT's static masking with dynamic masking, ensuring that the exact same tokens are not masked every time a sequence is fed to the model during training. Finally, they completely removed the next sentence prediction (NSP) objective, which was found to hurt downstream task performance. These optimizations resulted in RoBERTa outperforming BERT and achieving new state-of-the-art results across three major natural language understanding benchmarks: GLUE [26], RACE [13], and SQuAD [22].

Building directly upon RoBERTa's insights, Conneau et al. [2020] introduced XLM-R (Cross-lingual Language Model - RoBERTa). The authors specifically targeted the poor performance of previous multilingual models, like mBERT [10] and XLM [9], on low-resource languages. They argued that relying on Wikipedia for pretraining data left the models severely undertrained. As a result, XLM-R was trained on a massive 2.5-terabyte CommonCrawl corpus including over 100 languages, which increased the available data for low-resource languages. The authors also address the curse of multilinguality. This is the phenomenon where scaling a model to include more languages initially benefits low-resource languages through positive transfer, but eventually causes dilution that degrades overall performance across all languages. To represent 100 languages within a single model, the authors significantly increased the model's capacity and expanded the vocabulary size to 250,000 tokens. Furthermore, they tuned the language sampling rates to maintain a balance between high-resource and low-resource languages during training. XLM-R's results on multiple cross-lingual understanding benchmarks proved for the first time that a massively multilingual model could match the performance of strong monolingual models without sacrificing its cross-lingual capabilities. Because of its impressive performance in cross-lingual understanding, XLM-R serves as a foundational base model for some African language adapted models such as Nguni-XLMR [17] and AfroXLMR [3].

2.2 Encoder-Decoder Models

Encoder-only models were designed to produce a single prediction per input token or for an entire input sequence [21]. This made them highly effective in NLU tasks such as classification, but they were unsuited in NLG tasks such as translation. The field pivoted to encoder-decoder models when Raffel et al. [2020] introduced the Text-To-Text-Transfer-Transformer model (T5) which employed a unified text to text framework. This fundamentally involved considering NLP tasks (primarily NLG tasks) as consuming input text and generating output text, where the input text concatenates a task prefix to instruct the model on its task. Through extensive ablation studies, the authors proved encoder-decoder architecture consistently outperformed both decoder-only and prefix language models when applied to their text-to-text framework. Although T5 departs from BERT's encoder-only architecture for the original encoder-decoder architecture [25], they adapt BERT's training objective to introduce a span-corruption objective. Because this framework requires the model to generate sequence outputs, T5 departs from BERT's encoder-only architecture and instead utilizes the original encoder-decoder architecture [25]. Instead of predicting individual masked tokens, T5 randomly drops out 15 percent of the input tokens, replacing consecutive spans of dropped tokens with a single unique sentinel token. The decoder is then tasked with generating only the missing spans, significantly reducing the computational cost of pretraining.

To satisfy the massive data requirements of this model, the authors introduced the Colossal Clean Crawled Corpus (C4), consisting of 750GB of cleaned English web text from Common Crawl. Large amounts of data were needed because the authors discovered that repeating pretraining data on a scaled-down version of C4 caused the model to memorize the training data. Interestingly, their scaling experiments directly aligned with the core insights of RoBERTa [15]. They demonstrated that increasing dataset size, model parameters, and training time consistently yielded massive performance gains, solidifying that earlier base models were severely undertrained. Raffel et al. [2020] scaled T5 up to 11 billion parameters and achieved state-of-the-art results on 18 NLP benchmarks. However, because the C4 dataset was filtered to contain only English text, T5 failed to be considered a language-agnostic model as its results on translation tasks were subpar.

mT5 [29] is a multilingual variant of T5 that has been trained on 101 languages, including low-resource languages. It directly addresses the poor performance of T5 on multilingual tasks, such as translation, by creating mC4, a multilingual extension of the C4 corpus which covers 101 languages. In order to show that the T5 framework was transferable to multilingual settings, mT5 followed the same architecture and training objective as T5. When exploring model scaling, the authors encountered a fascinating trend in regards to the curse of multilinguality. mT5-Small and mT5-Base performed worse than their English-only T5 counterparts on the SQuAD benchmark for reading comprehension, but the larger mT5 models performed on par. This aligns with the findings of XLM-R, where researchers increased the model's size to effectively handle 100 languages. Furthermore, they showed that as model capacity increases, the performance gap between zero-shot transfer and

translate-train decreases. Since mT5 is a generative model, it encountered accidental translation, where the model would retrieve the correct answer in the target language before erroneously translating it into English. This was a unique challenge not faced by encoder-only models like XLM-R. The authors hypothesized that this happens because, during supervised finetuning, the model's weights bias towards English. Using domain-preserving training (adding unlabeled multilingual pretraining data during finetuning) reduced the rate of accidental translations, however, it did not completely eradicate it. mT5 proved that a single, large encoder-decoder architecture could be adapted to multiple languages and excel.

2.3 Decoder-only Models

Although the earliest decoder-only models, such as GPT-1 [20], were previously criticized by Devlin et al. [2019] for their purely left-to-right attention mechanism, the decoder-only architecture has seen a resurgence in recent years due to the development of highly scaled and efficient models. A recent decoder-only model is LLaMA [24], which shifted the research focus from increasing model parameters to prioritizing training data amount. Touvron et al. [2023] demonstrated that training smaller decoder models on significantly more data achieves state-of-the-art performance that rivals much larger models. What makes LLaMA appealing is that it was trained exclusively on publicly available datasets and designed to be extremely efficient during inference. Because of its open-source nature, it has been widely utilized in African NLP research, serving as a foundational base model for adapted architectures such as Luga-LLaMA [5]. Luga-LLaMA was adapted to 16 low-resource African languages and proved that scaled decoders can be effective in low-resource NLP tasks. Subsequently, the authors demonstrated that combining curated African language data with high-quality English educational documents improves the model's performance on downstream tasks.

3 Language Adaptation Methods

3.1 Transfer Learning and Cross-lingual Transfer

Training large language models from scratch requires enormous amounts of computational resources and massive, high-quality datasets, which are unavailable for many low-resource African languages. As a result, NLP researchers rely on transfer learning in the absence of large data. Transfer learning is a paradigm where a model is pretrained on a data-rich task and then fine-tuned on a downstream task using a smaller dataset [21]. While models like T5 [21] successfully explored transfer learning on an English text-to-text framework, researchers also extend this concept to low-resource languages. Cross-lingual transfer is the ability of a model to leverage knowledge it has learned in a high-resource language and apply it to a low-resource language. It can also apply to similarly resourced languages sharing knowledge between each other. The most extreme case of this is zero-shot cross-lingual transfer, where a multilingual model is fine-tuned on a specific task in one language (e.g. English) and then directly evaluated on a target language, even though it never saw supervised examples in that target language. By pretraining a model on a mixture of diverse languages, the model's internal representations encode

shared knowledge and information across languages. A language-agnostic model is one where these internal representations are aligned, meaning the model processes concepts similarly regardless of the specific input language. Multilingual models have proven their potential, but they still suffer from poor African language representation. For instance, the pretraining corpora for XLM-R [8] and mT5 [29] cover only 8 and 13 African languages, respectively. When an African language is completely unseen during the initial pretraining phase, cross-lingual transfer is ineffective, resulting in a huge drop in downstream task performance [3]. We explore specific adaptation methods to extend foundational models to new languages.

3.2 Continued Pretraining

Continued pretraining (CPT) aims to adapt an already pretrained model to a new language or domain. To adapt existing pretrained models to African languages, researchers frequently use monolingual CPT, a form of continued pretraining where the model undergoes further unsupervised training on a monolingual corpus of the target language. Adelani et al. [2021] demonstrated this process during the development of their MasakhaNER benchmark, where they adapted mBERT and XLM-R to specific languages, such as Amharic and Yorùbá. They found that applying monolingual CPT and replacing mBERT's vocabulary was essential for unseen writing systems like the Amharic script, which resulted in a massive 60 percent performance improvement over the base mBERT model. They showed this approach yielded a 1 to 7 percent improvement in named entity recognition (NER) F1 scores over the baseline XLM-R and mBERT models.

Despite the success of monolingual CPT, which produces strong results for African languages, Alabi et al. [2022] argued that adapting a model to each target language requires a significant amount of disk space and limits cross-lingual transfer since each model would specialize in only one language. Multilingual CPT was introduced to resolve these issues. Instead of continually pretraining a model on a single language, this approach adapts a model on a massive combined corpus of different languages simultaneously. By applying multilingual CPT to XLM-R using 17 diverse African languages and 3 high-resource languages spoken on the African continent, they created AfroXLMR. This model achieved competitive and sometimes superior F1 scores across NER, sentiment analysis, and news classification against individual XLM-R models that were trained on only one language. Additionally, Alabi et al. [2022] reduced the vocabulary of AfroXLMR to manage the massive size of multilingual models. Multilingual CPT proved itself to be highly performant and resource-efficient for low-resource settings.

Following the success of multilingual CPT on encoder-only models, Adelani et al. [2022] demonstrated that this adaptation method is equally viable for encoder-decoder architectures. By applying continued pretraining on a corpus of 17 African languages, the authors adapted mT5 and ByT5 to create AfriMT5 and AfriByT5, respectively. To evaluate these models, they introduced an African news translation benchmark: MAFAND-MT. A comparison between the base models and their adapted counterparts revealed that the adapted models yielded better results, however, these results were relatively modest. For instance, AfriByT5 improved upon the base

Table 1: Summary of African-Adapted Language Models

Model Name	Authors & Paper	Model Type	Parameters	African Languages Covered
AfroXLMR	Alabi et al. [2022] in "Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning" [3]	Encoder-only	117M (Mini) 270M (Base) 550M (Large)	17 Languages: Afrikaans, Amharic, Chichewa, Hausa, Igbo, Kinyarwanda, Kirundi, Malagasy, Naija, Oromo, Sesotho, Shona, Somali, Swahili, isiXhosa, Yorùbá, isiZulu.
AfriMT5	Adelani et al. [2022] in "A Few Thousand Translations Go A Long Way! Leveraging Pre-trained Models for African News Translation" [2]	Encoder-Decoder	580M	17 Languages: Same 17 languages as AfroXLMR.
AfriByT5	Adelani et al. [2022] in "A Few Thousand Translations Go A Long Way! Leveraging Pre-trained Models for African News Translation" [2]	Encoder-Decoder	580M	17 Languages: Same 17 languages as AfroXLMR.
AfriMBART	Adelani et al. [2022] in "A Few Thousand Translations Go A Long Way! Leveraging Pre-trained Models for African News Translation" [2]	Encoder-Decoder	610M	17 Languages: Same 17 languages as AfroXLMR.
Lugha-LLaMA	Buzaaba et al. [2025] in "Lugha-Llama: Adapting Large Language Models for African Languages" [5]	Decoder-only	8B	16 Languages: Afrikaans, Amharic, Hausa, Igbo, Kinyarwanda, Malagasy, Nyanja, Oromo, Shona, Somali, Sotho, Swahili, Tigrinya, Xhosa, Yoruba, Zulu.
Nguni-XLMR	Meyer et al. [2024] in "NGLUEni: Benchmarking and Adapting Pretrained Language Models for Nguni Languages" [17]	Encoder-only	355M (Large)	4 Languages: isiXhosa, isiZulu, isiNdebele, Siswati.
Nguni-ByT5	Meyer et al. [2024] in "NGLUEni: Benchmarking and Adapting Pretrained Language Models for Nguni Languages" [17]	Encoder-Decoder	1.23B (Large)	4 Languages: isiXhosa, isiZulu, isiNdebele, Siswati.

ByT5 model by only 0.6 to 0.9 BLEU points in English or French to target language and target language to English or French translations, respectively. While multilingual CPT provides linguistic foundation and further augmentation, the effective strategy is the supervised fine-tuning stage. Adelani et al. [2022] proved that fine-tuning on a small, high-quality dataset enables these models to achieve highly competitive performance.

We note that this approach tends to use dozens of linguistically diverse languages for African models, Meyer et al. [2024] approaches multilingual CPT with an efficiency mindset similar to Adelani et al. [2022]. They demonstrated that continued pretraining is significantly more effective when restricted to a smaller group of languages that are linguistically similar. Having introduced the NGLUEni benchmark for the South African Nguni languages (isiXhosa, isiZulu, isiNdebele, and Siswati), they applied multilingual CPT to XLM-R [8] and ByT5 [28], creating Nguni-XLMR and Nguni-ByT5. To do this, they mixed larger monolingual corpora for isiZulu and isiXhosa with smaller, resource-starved languages: isiNdebele and Siswati. Nguni-XLMR consistently outperformed Afro-XLMR

across NLU tasks, with Meyer et al. [2024] arguing that using four highly related languages allows for better cross-lingual transfer and avoids negative interference from other languages. To test zero-shot cross-lingual transfer, they fine-tuned the models on training data from isiXhosa and then evaluated them on the other three Nguni languages. The resulting Nguni-XLMR-large model achieved an average zero-shot performance improvement of over 20 percent for NLU tasks in isiNdebele and Siswati compared to the base XLM-R model. For NLG, Meyer et al. [2024] chose ByT5 [28] because it processes text byte-by-byte rather than using subwords, which is advantageous for agglutinative languages. Across the generative tasks, Nguni-ByT5-large outperformed AfriByT5. They also designed a zero-shot cross-domain test where they trained Nguni-ByT5 to write headlines for news in isiXhosa and then tested it on government documents across all Nguni languages. Nguni-ByT5 achieved significantly better cross-lingual transfer to the data-starved isiNdebele and Siswati languages than the broadly adapted AfriByT5 [17].

4 Learning Dynamics

Language adaptation methods have proven effective, yet existing evaluations typically assess only final performance, leaving open the question of how and when linguistic knowledge is acquired during training. Learning dynamics involves analysing a model at multiple checkpoints throughout its training to understand how and when its linguistic capabilities emerge. Prior work in learning dynamics groups methods into four categories.

4.1 Vector Analysis

Vector analysis is an approach that examines the model's hidden layers using mathematical tools to measure how these representations shift over time. Pioneering this approach, Saphra and Lopez [2019] delved into the learning dynamics of LSTMs. They used SVCCA (singular vector canonical correlation analysis), a mathematical method of comparing representations from a training model with the representations of already specialised tagging models. This allowed them to inspect the geometry of the training model's hidden layers without needing annotated data. They discovered that LSTMs master basic grammar rules, like part-of-speech (POS), very early in training, while more complex concepts such as topic require a longer time to emerge because the task depends on longer-range context. Müller-Eberstein et al. [2023] expanded upon vector analysis by evaluating BERT and comparing subspaces. Subspaces are regions within a model that are dedicated to specific linguistic tasks. They calculated principal subspace angles (SSAs) and the center of gravity (COG), allowing them to compare task-specific subspaces and find at what layer depth they can find the most important task-relevant information. Their vector analysis found that there is a critical phase between 1,000 and 10,000 steps. Furthermore, they found that in early training, subspaces dip into the lower layers and share information, before specialising and disentangling much later in training.

4.2 Structural Probing

Structural probing involves attaching a small classifier to the model's layers. This classifier is trained on the model's internal representations at that layer while keeping the original model's parameters frozen. This helps to identify the training step and layer that the model learns language rules like part-of-speech.

Chiang et al. [2020] used a modified edge probing framework on ALBERT to investigate its learning dynamics. They extracted several checkpoints over 1 million pretraining steps and attached linear classifier probes to the model's layers. They found that syntactic and semantic knowledge develop simultaneously during the first 100,000 steps, with syntactic performance improving rapidly. Furthermore, semantic tasks like semantic role labeling (SRL) peaked at around 150,000 steps and then started to decay slightly afterwards.

Building on this, Liu et al. [2021] conducted a study on RoBERTa using the structural probing of the linguistic knowledge and transferability (LKT) suite [14]. Similar to Chiang et al. [2020], they trained linear classifiers on RoBERTa's frozen representations. They found that linguistic and syntactic knowledge is acquired extremely fast. Within the first 20 percent of total pretraining steps, 97 percent of overall performance had been reached. These results align with

Chiang et al.'s [2020] findings that syntactic knowledge is learned rapidly and early in training.

4.3 Zero-shot

Zero-shot evaluation, also known as behavioral probing, assesses a model's natural behavior on a task at different stages of training. It differs from structural probing, which attaches a classifier to the model's internal layers, in that we simply test the model's inherent capability at a given checkpoint.

Chiang et al. [2020] analysed how the model predicted words with different parts of speech through zero-shot masked language modelling prediction and reconstruction. They found that vocabulary is learned at different points in time. Function words, such as determiners and prepositions, are learned first, while content words, such as proper nouns, take longer to emerge.

Liu et al. [2021] utilized a wide range of behavioral probes to test RoBERTa's natural behavior. They used several benchmarks for linguistic knowledge, factual knowledge, common sense, and reasoning. Their tests confirmed linguistic knowledge is learned rapidly, which supported their LKT probes. Common sense and factual knowledge are acquired much slower, with factual knowledge being more sensitive to the dataset and reasoning being learned unstably.

Further research by Choshen et al. [2022] relied exclusively on zero-shot evaluations to study the learning dynamics of several models with different architectures, such as decoder-only transformers, Transformer-XL, and LSTMs. Using the BLiMP [27] dataset, they tested the models inherent ability to assign higher probabilities to grammatically correct sentences over incorrect ones. They discovered that there is a consistency to learning, where the order a model learns grammar is consistent, regardless of its initialisation, architecture, or training data. While a larger model might learn faster, the learning trajectories between models remained consistent. Choshen et al. [2022] align their research with the findings of Saphra and Lopez [2019] on the observation that models slowly learn to rely on context later in training, and they reinforce Liu et al.'s [2021] work, which saw that basic linguistic and syntax rules are acquired early in training.

Building directly upon Choshen et al.'s [2022] work, Evanson et al. [2023] investigated whether language models follow similar learning trajectories to human children. They pretrained 48 GPT-2 models from scratch and used zero-shot evaluations on several benchmarks. Their findings strongly support Choshen et al. [2022], confirming that models acquire linguistic skills in a highly predictable and consistent order. They found that models do not learn sequentially but rather in parallel. Simple and complex linguistic skills develop simultaneously from the start of pretraining, but progress at different speeds. Some learning stages are shared between children and language models. For three specific syntactic tests (subject-verb agreement, wh-questions, and relative clauses), the majority of the models learned them in the exact same order as human children. However, Evanson et al. [2023] also stated that models can achieve high scores using superficial statistical heuristics rather than actually learning the grammar.

Biderman et al. [2023] performed evaluations on their novel Pythia suite, which comprises of decoder-only models ranging from

70 million to 12 billion parameters. They discovered that a model’s ability to memorise the text it trains on is purely random. Text is memorised uniformly. They also examined whether a model’s accuracy improves if a specific fact appears frequently in the training data. High term frequency was found to be emergent. It only begins to correlate with accuracy later in training (at about 65,000 steps) and is exclusively observed in models larger than 2.8 billion parameters. Models below this size struggle to accurately recall facts no matter how many times they appear in the training data.

Recent research by Elhady et al. [2025] explores how these learning dynamics apply to continued pretraining for language adaptation. They adapted English-centric decoder-only large language models, to Basque, Arabic, and Indonesian. They used validation perplexity and in-context learning as their evaluation metrics for each of their intermediate checkpoints. To evaluate ICL in a language-agnostic manner, they introduced Copain. This benchmark requires models to guess mathematical or alphabetical patterns with no instructions. They compared models trained exclusively on the target language versus models trained with a mixture that contained 20 percent English data. They discovered that the perplexity of both types of models was similar, however, the models trained exclusively on the target language had severely poor performance on downstream tasks. They found that early in training, a model experiences a strong shift in its parameters (a critical period) causing catastrophic forgetting of prior knowledge if English is not included in the training data. Including English data, using curriculum learning or exponential moving average prevents catastrophic forgetting.

4.4 Downstream Fine-Tuning

Downstream task analysis takes a model at various checkpoints in the training process and fully fine-tunes it on a downstream task to evaluate when the model gains enough knowledge to complete that task.

Chiang et al. [2020] evaluated ALBERT by fine-tuning its checkpoints and evaluating on the GLUE [26] benchmark. During their downstream task analysis, they discovered that a model’s performance does not increase much after reaching the 250,000 step checkpoint. This proved that early checkpoints could be fine-tuned and still achieve competitive performance with models that trained for much longer. Most notably, Chiang et al. [2020] argue that linguistic knowledge and world knowledge do not generally improve as pretraining continues.

Liu et al. [2021] fine-tuned 14 intermediate checkpoints for RoBERTa on 8 different benchmarks. Their results aligned with their structural probes and zero-shot evaluations. For instance, single-sentence tasks achieved 97 percent of their best performance within 16-32 percent of the total pretraining time. Tasks that tested commonsense and complex knowledge required longer pretraining times. Most importantly, the authors challenge Chiang et al. [2020] claim that a model’s world knowledge and linguistic knowledge does not improve as pretraining continues, finding that it in fact does improve.

During their fine-tuning experiments, Müller-Eberstein et al. [2023] took checkpoints at intervals of 0, 1,000, 5,000, 20,000, and 2 million steps and fully fine-tuned them on different downstream

tasks. These results aligned with their probing insights. The greatest improvement to downstream performance occurred between 1,000 and 20,000 steps of pretraining. They discovered at around 10,000 step the encoder becomes beneficial to NLP tasks. At 5,000 steps, the model’s performance was suboptimal. More complex tasks like question answering only acquired usable performance after 20,000 steps.

5 Critical Comparison of Prior Work

The field of NLP has seen the emergence of different pretrained architectures. Initially, encoder-only models such as BERT [10] and XLM-R [8] set a standard for how models should perform on NLU tasks. Then, encoder-decoder models such as T5 [21] and mT5 [29] combined NLU and NLG (with a primary focus on NLG) tasks into a single framework and showed impressive performance at generation. Recently, large-scale decoder-only models like LLaMA [24] have become a focal point of the field. However, these models and architectures all have something in common: they require vast amounts of data for pretraining. As a result, foundational models are often pretrained on high-resource languages, leaving low-resource languages to find other adaptation techniques to improve performance and close the gap between them.

Continued pretraining aims to improve the performance of low-resource languages by leveraging an already existing PLM’s knowledge. Monolingual CPT adapts a model to a single language only. This approach has been shown to improve performance for African languages, especially for unseen writing systems [1]. However, this creates highly specialized, performant models that require substantial disk space for each new target language and limits cross-lingual transfer. Multilingual CPT approaches language adaptation from an efficiency standpoint. By adapting a single model to multiple languages simultaneously, cross-lingual transfer improves and less disk space is needed. Performance of multilingual CPT models has been shown to be competitive with monolingual CPT models [3]. However, these models can experience the curse of multilinguality, where performance across all languages drops because the model’s capacity experiences dilution [8]. Furthermore, using a smaller group of linguistically similar languages improves cross-lingual transfer and avoids interference from other distinct languages [17]. Although we know the tradeoffs of monolingual and multilingual CPT, the current literature on CPT for low-resource languages evaluates performance based only at the end state of training. While we know these adaptation methods work, the actual underlying understanding of how they work is cast aside.

The field of learning dynamics seeks to understand exactly when and how models acquire linguistic capabilities. Research on learning dynamics, across different architectures, reveals a hierarchical timeline of learning. In the earliest stages of training, models acquire basic syntactic rules rapidly [16, 18, 23]. Once syntax is established, the model enters an intermediate phase where it slowly acquires facts and world knowledge [4, 16, 18]. Finally, reasoning and deep semantics emerge last, as the model requires longer training to develop long-range contextualisation [18]. This learning trajectory is consistent regardless of the model’s size or initialisation [7]. These linguistic skills develop in parallel and align with human linguistic development [12].

These learning dynamics are derived primarily from pretraining models. CPT, on the other hand, is a completely different learning process on the training pipeline. A model that is pretrained from scratch starts with randomly initialised weights and acquires knowledge on syntax, facts, and reasoning from a blank slate. When we apply CPT to a model, it already has internal representations and world knowledge, therefore, the model must align its learned representations to new language vocabulary. This means it is unknown if the hierarchical timeline of learning is consistent during CPT. CPT has existed throughout the different model architectures, however, learning dynamics related to it are not explored sufficiently. Elhady et al. [2025] have explored learning dynamics on CPT in new languages, however, they restrict themselves to decoder-only models and ICL as evaluation metrics. The adaptation of encoder-only and encoder-decoder models has not been explored. Furthermore, Müller-Eberstein et al. [2023] identified a critical phase between 1,000 and 10,000 steps (making up just 0.5 percent of total training time) where models saw their most significant performance gains. They observed that performance for most tasks reached 90 to 95 percent at 20,000 steps. If this finding holds true for CPT, future researchers would not have to run CPT for unnecessarily long steps. Therefore, we investigate the learning dynamics of CPT by extracting intermediate checkpoints of encoder-only and encoder-decoder models and fine-tuning them on downstream tasks. This allows us to investigate when usable linguistic knowledge emerges during the adaptation of South African languages and if the dynamics of pretrained models align with CPT.

6 Conclusion

We explored the different transformer architectures, language adaptation techniques, and the learning dynamics of pretrained language models. While model architectures have evolved, their reliance on large-scale datasets causes inclusivity issues for low-resource languages. Continued pretraining is used by researchers to adapt pretrained models to new low-resource languages, with the monolingual variant offering high performance and the multilingual variant offering efficiency. Learning dynamics seeks to understand how and when models acquire linguistic knowledge. Much of the current research focuses on the dynamics during pretraining. A timeline of what models learn first during pretraining has been established: namely syntax, then facts and common sense, and lastly reasoning. However, hardly any research has been done on knowledge acquisition during CPT, despite its widespread use. We plan to investigate the learning dynamics of encoder-only and encoder-decoder models as they adapt to South African languages through post-training. By extracting intermediate checkpoints during CPT and finetuning them on downstream tasks, we aim to compare the established learning dynamics of pretraining with the dynamics that occur during CPT.

References

- [1] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics* 9 (2021), 1116–1131.
- [2] David Ifeoluwa Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiters, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. 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*. 3053–3070.
- [3] Jesujoba 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*. 4336–4349.
- [4] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*. PMLR, 2397–2430.
- [5] Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025. Lughu-Llama: Adapting Large Language Models for African Languages. *arXiv preprint arXiv:2504.06536* (2025).
- [6] 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)*. 6813–6828.
- [7] Leshem Choshen, Guy Hacohen, Daphna Weinsshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8281–8297.
- [8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 8440–8451.
- [9] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems* 32 (2019).
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.
- [11] Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. Emergent abilities of large language models under continued pre-training for language adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 32174–32186.
- [12] Linnea Evanson, Yair Lakretz, and Jean-Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages?. In *Findings of the Association for Computational Linguistics: ACL 2023*. 12205–12218.
- [13] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 785–794.
- [14] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1073–1094.
- [15] 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. *arXiv preprint arXiv:1907.11692* (2019).
- [16] Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does RoBERTa know and when?. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 820–842.
- [17] Francois Meyer, Haiyue Song, Abhisek Chakrabarty, 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)*. 12247–12258.
- [18] 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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 13190–13208.
- [19] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. doi:10.18653/v1/N18-1202
- [20] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

- [22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2383–2392.
- [23] Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3257–3267.
- [24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [26] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*. 353–355.
- [27] 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.
- [28] 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 (2022), 291–306.
- [29] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 483–498.