

Continued Pretraining Learning Dynamics of Encoder-Only Models on South African Languages

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

ABSTRACT

Continued pretraining (CPT) is widely used to adapt pretrained language models (PLMs) to low-resource languages. However, most studies evaluate models only after training and provide limited insight into the underlying adaptation process. Studying learning dynamics addresses this limitation by analysing how model performance and internal representations change throughout training. However, learning dynamics remain unexplored for CPT-based language adaptation. To address this gap, we investigate the learning dynamics of four encoder-only PLMs during monolingual CPT adaptation to isiXhosa and isiZulu by analysing how their downstream performance and cross-lingual representation geometry change across CPT checkpoints. We find that downstream performance improves most rapidly during early CPT before slowing and largely stabilising. Our geometric analysis shows that CPT can increase English–target-language alignment, with stronger alignment and more effective use of representational capacity generally occurring in intermediate and upper-middle hidden layers before representational degeneration becomes more pronounced towards the final layers. Finally, we find that models with less prior exposure to the target languages generally show larger downstream and geometric changes, while models with more prior exposure change less and may even degrade under additional CPT. Our study provides the first insights into how encoder-only models adapt to African languages during CPT, the most established and reliable approach for developing state-of-the-art PLMs for African languages. In doing so, we extend learning-dynamics research beyond pretraining from scratch to help inform more effective adaptation of PLMs to African and other low-resource languages.

KEYWORDS

Natural Language Processing, Learning Dynamics, Large Language Models, Continued Pretraining, AI Interpretability, Low-Resource Languages, African Languages, Natural Language Understanding

1 INTRODUCTION

Pretrained language models (PLMs) have become the foundation of modern natural language processing (NLP), achieving strong performance across a wide range of language understanding [9, 11] and generation tasks [43, 51]. These Transformer-based models contain millions of parameters whose interactions are difficult to interpret [52], making it difficult to determine how they represent linguistic knowledge and when this knowledge develops during training [5, 35]. As a result, PLMs remain largely black boxes, which limits our understanding of how their language representations develop during initial pretraining and subsequent adaptation to new domains, languages, and tasks.

Following initial pretraining, PLMs can be adapted to new languages through continued pretraining (CPT) [4, 19]. CPT involves further training a PLM on additional unlabelled data from a target language, which allows the model to transfer knowledge and abilities learned during pretraining to the target language [14]. Multilingual PLMs learn useful linguistic knowledge from their training languages, which provide a starting point for adaptation to a specific target language [4, 14, 41]. Compared with pretraining from scratch, CPT is more efficient [32], as training PLMs from scratch is resource-intensive in terms of data and compute [4]. CPT has been successfully applied to develop state-of-the-art PLMs for a range of African languages, consistently improving their downstream performance on tasks such as text classification [2, 4] and machine translation [32, 57]. However, despite its effectiveness and widespread use in adapting PLMs to low-resource languages, it remains largely unclear *why* CPT works, *when* it is most effective, and *where* it fails.

One way to better understand how PLMs adapt to new languages is to study their learning dynamics. Most research evaluates CPT-adapted models only after training is complete [1, 4, 32, 57], which provides limited insight into how their performance and representations change during adaptation. Studying a model’s *learning dynamics* addresses this limitation by analysing its internal representations and performance at intermediate checkpoints throughout training [35, 47]. However, most existing learning dynamics research has analysed models pretrained from scratch rather than models adapted to new languages through CPT [8, 16]. CPT begins with a PLM that already contains knowledge and representations from pretraining [4, 19], while pretraining starts from a randomly initialised model that must develop these representations from scratch [7, 47]. Because these processes differ fundamentally, it remains unclear whether learning patterns seen during pretraining also appear during CPT. Although recent work has investigated learning dynamics during CPT [14], this work does not include any low-resource African languages. The learning dynamics of CPT for low-resource South African languages therefore remain unexplored.

To address this gap, we analyse the learning dynamics of encoder-only PLMs adapted through CPT to isiXhosa and isiZulu, two South African Nguni languages. We also examine how closely the observed learning dynamics reflect those seen during pretraining from scratch. Specifically, we address three research questions:

1. How does the downstream task performance of encoder-only models change during CPT adaptation to isiXhosa and isiZulu?
2. How do cross-lingual alignment and isotropy change across the layers of encoder-only models during CPT adaptation to isiXhosa and isiZulu?

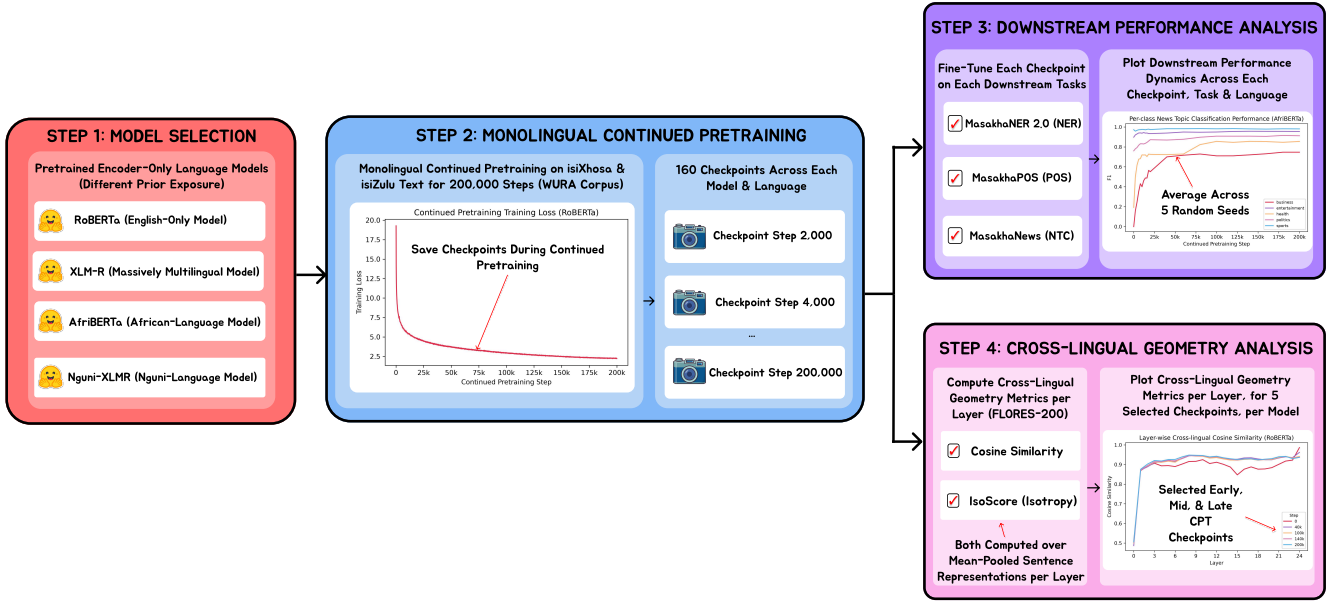


Figure 1: Overview of our methodology. We continue pretraining four encoder-only PLMs with different levels of prior exposure to isiXhosa and isiZulu, evaluate their checkpoints on three downstream NLU tasks, and analyse the cross-lingual geometry of selected checkpoints using layer-wise cosine similarity and isotropy.

- How does a model’s prior exposure to the target languages affect its learning dynamics during CPT?

We adopt an interpretability-focused approach, combining downstream evaluation with analysis of internal representations to investigate these questions. To do so, we continue pretraining XLM-R [9], RoBERTa [27], AfriBERTa [36], and Nguni-XLMR [32] monolingually on isiXhosa and isiZulu text and extract 20 checkpoints throughout training. We study their learning dynamics using two complementary methods. First, we fine-tune and evaluate these checkpoints on named entity recognition (NER), news-topic classification (NTC), and part-of-speech (POS) tagging and analyse how their downstream performance changes throughout CPT. We then analyse the geometry of cross-lingual representations at selected checkpoints to understand how cross-lingual alignment and representation-space utilisation change during CPT. We examine this geometry using two metrics: *cosine similarity*, which measures alignment between representation vectors in a shared representation space [23], and *isotropy*, which measures how evenly representations are distributed across the space [46].

We find that downstream performance improves most rapidly during early CPT before slowing and largely stabilising, with NTC generally showing larger improvements than NER and POS tagging. We further find that CPT changes cross-lingual geometry in two ways: it can bring the representations of English and the target languages closer together, while also improving isotropy so that the joint representation space makes more effective use of its available capacity. We also find clear layer-wise differences, as alignment and isotropy are generally strongest in the intermediate and upper-middle layers before representational degeneration becomes more

pronounced in the final layers. Finally, models with less prior exposure to isiXhosa and isiZulu generally show larger downstream and geometric changes, while models with more exposure change less and may even degrade under additional CPT.

Our study is the first to provide insights into how encoder-only models adapt to African languages during CPT, the most established and reliable approach for developing state-of-the-art PLMs for African languages [1, 4, 32, 57]. In doing so, we extend learning-dynamics research beyond pretraining from scratch to help inform more effective adaptation of PLMs to African and other low-resource languages.

2 BACKGROUND

We begin by reviewing the concepts and literature relevant to our study. We first introduce encoder-only PLMs and CPT as a technique for adapting PLMs to low-resource languages. We then discuss learning dynamics and the geometry of cross-lingual representations as approaches for understanding how models learn and adapt to new languages.

2.1 Encoder-Only Models

Encoder-only models are a class of pretrained transformers used for natural language understanding (NLU) tasks, which involve understanding and analysing the meaning of input text [11, 13]. They are trained using a masked language modelling (MLM) objective, where a proportion of input tokens are randomly masked, and the model learns to predict the masked tokens from the surrounding context in both directions [11]. This bidirectional objective enables encoder-only models to learn rich contextual representations from

input sequences, which makes them well-suited for language understanding tasks such as NER [3, 41] and POS tagging [12, 32]. While generative decoder-only models are widely used in modern NLP and achieve strong performance on high-resource benchmarks [49, 56], specialised encoder-only models remain widely used and can outperform decoder-only models in low-resource settings on tasks such as NER, POS tagging, and sentiment analysis [28, 37].

BERT [11] established the MLM pretraining paradigm, while RoBERTa [27] improved BERT’s pretraining strategy. XLM-R [9] extended RoBERTa to 100 languages, enabling shared multilingual representations and cross-lingual transfer. We focus on this model class to observe how language understanding develops as PLMs adapt to low-resource languages through CPT.

2.2 Continued Pretraining for Low-Resource Language Adaptation

CPT is a technique for adapting PLMs to new languages by further training them on additional target-language data using their original pretraining objective [4, 19]. It relies on transfer learning, which allows knowledge learned during pretraining to support adaptation to new contexts [19, 21] and languages [9, 41]. CPT is more efficient than training from scratch because it starts with a model that already contains representations developed during pretraining [19, 21], which reduces the data and compute required for adaptation [4, 32]. This efficiency makes CPT useful for adapting PLMs to low-resource languages, where limited data make training models from scratch difficult [1, 32].

CPT can be applied monolingually to a single target language or multilingually to multiple languages simultaneously [4]. Multilingual CPT can be more resource-efficient because a single model adapts to multiple languages, while also benefiting from cross-lingual transfer between similar languages in the training data [4, 32]. Examples of PLMs adapted to African languages through CPT include Afro-XLMR [4] and Nguni-XLMR [32], which extend XLM-R and include South African languages such as isiXhosa and isiZulu.

However, the effectiveness of multilingual CPT depends on the linguistic relationships between the languages included during training. Including too many unrelated languages can reduce a model’s capacity for individual languages, which is known as the curse of multilinguality [9], while restricting CPT to linguistically similar languages can improve cross-lingual transfer and reduce negative interference [32]. CPT’s effectiveness also depends on a model’s prior exposure to a target language [4, 14, 32], as models with greater exposure already encode more information about that language and may require less additional adaptation.

2.3 Learning Dynamics

Learning dynamics describe how a model’s behaviour and internal representations change throughout training. Rather than evaluating a model only after training is complete, researchers study learning dynamics by analysing a model’s internal representations and downstream performance at intermediate training checkpoints [7, 47]. Approaches for analysing learning dynamics include probing, which evaluates information encoded in frozen intermediate representations using lightweight classifiers [7, 26]; internal representation analysis, which directly analyses model parameters

or hidden representations [35, 47]; and downstream evaluation, which measures model performance on downstream tasks either through zero-shot evaluation without any fine-tuning [6, 8, 14, 16] or by fine-tuning intermediate checkpoints on those tasks [30, 35].

These approaches have identified broad learning patterns during pretraining, where basic syntactic and semantic knowledge develops early and more complex factual, commonsense, and reasoning abilities develop later [7, 26, 35]. While learning dynamics during pretraining from scratch have been studied extensively, they remain unexplored in the context of CPT for low-resource African language adaptation. Recent work by Elhady et al. [2025] investigates the learning dynamics of decoder-only models adapted to Basque, Arabic, and Indonesian via CPT, where they identify a critical period in the early stages of CPT during which the models experience catastrophic forgetting of prior in-context learning abilities. However, this work does not include any encoder-only models or low-resource African languages, which motivates our study of the learning dynamics of encoder-only models during CPT adaptation to South African languages.

2.4 Geometry of Cross-Lingual Representations

Multilingual PLMs learn representations for multiple languages within a shared representation space [41]. Cross-lingual alignment refers to the similarity of multilingual representations across languages in this shared space [23]. These representations can be static embeddings [33], which assign a single vector to each word or token, or contextualised embeddings, whose representations vary with the surrounding context [11, 39]. Hämmerl et al. [2024] distinguish between geometric alignment, where semantically similar representations across languages lie closer together, and functional alignment, where a prediction head trained on a source language can accurately classify target-language inputs. Greater alignment between low-resource and high-resource languages such as English can support cross-lingual transfer and is associated with improved downstream performance [24]. We analyse geometric alignment using cosine similarity [23] to investigate whether isiXhosa and isiZulu become more closely aligned with English during CPT.

Representation spaces can also suffer from representational degeneration, where embeddings fail to use the full capacity of the available space [18, 31]. One form of degeneration is anisotropy, where representations concentrate in a narrow region of the embedding space [15, 45]. This can reduce the effective dimensionality of the representation space and has been associated with poorer model performance [18, 31, 46]. Anisotropy can also cause unrelated representations to have high cosine similarity, meaning that high similarity does not necessarily indicate meaningful alignment [17, 22]. We therefore measure isotropy alongside cosine similarity to assess whether CPT reduces representational degeneration and improves use of the representation space. Cosine similarity captures local, relational alignment between representations, while isotropy measures global utilisation of the representation space [31].

3 METHODOLOGY

We study the learning dynamics of encoder-only PLMs as they adapt to isiXhosa and isiZulu through monolingual CPT. To obtain the checkpoints used in our analyses, we continue pretraining four

encoder-only PLMs with different levels of prior exposure to isiXhosa and isiZulu and save intermediate checkpoints throughout CPT. We then analyse these checkpoints using two complementary methods. First, we analyse how downstream performance changes during CPT by fine-tuning model checkpoints on downstream NLU tasks. Second, we analyse the geometry of their cross-lingual representations during CPT using cosine similarity and isotropy to provide a model-internal perspective on the patterns observed during downstream performance analysis. In this section, we describe the CPT procedure and the methods used for our downstream performance and geometric analyses.¹ Figure 1 provides an overview of our CPT, downstream performance, and geometric analysis methodology, and Tables A1 and A2 (see Appendix A) provide our full CPT and fine-tuning hyperparameter configurations, respectively.

3.1 Target Languages & Continued Pretraining Data Preparation

isiXhosa and isiZulu are two South African languages in the Nguni subgroup of the Bantu language family, along with isiNdebele and Siswati [32]. Together, isiXhosa and isiZulu have approximately 24 million home language speakers [48] (see Table 1 for language statistics), representing approximately 40% of the South African population. Both languages are considered low-resource in NLP and have rich morphological structure [32, 42]. Studying both languages allows us to examine whether learning dynamics during CPT adaptation are consistent across closely related languages.

Table 1: Home language speaker counts and WURA corpus sizes for isiXhosa and isiZulu.

Language	isiXhosa	isiZulu
Speaker statistics		
Home language	9,786,928	14,613,202
WURA corpus size (words)		
Training split	13,104,825	31,584,011
Validation split	1,493,438	3,441,640

CPT Corpus. To study these dynamics, we require a suitable CPT corpus. We use the WURA corpus [38] as the source of our isiXhosa and isiZulu pretraining data, given its established use in prior pretraining work for these languages [28] and in studies of isiXhosa pretraining dynamics [30]. WURA is a curated multilingual corpus covering 16 African languages, created by filtering mC4 [54] to remove low-quality text and crawling additional verified news sources. The resulting corpus is cleaner and more informative than raw web-crawled text [25], improving training effectiveness by providing higher-quality data to our models [20, 44]. The corpus consists primarily of news articles, with each entry containing a headline and content field. The WURA isiXhosa and isiZulu training splits contain approximately 13 million and 31 million words, respectively (see Table 1 for full corpus statistics). We use each language’s training split for CPT and its validation split to evaluate

the model on unseen data during training. These splits provide sufficient data for our experiments while remaining computationally manageable.

Corpus Preparation. To prepare the corpus for CPT, we follow standard MLM pretraining practices [11, 27]. We first concatenate each article’s headline and content into a contiguous text sequence. We then tokenise the resulting texts using each model’s native tokeniser and concatenate the tokenised sequences across the corpus into a single sequence. Finally, we split the resulting sequence into fixed-length blocks of 512 tokens and discard any remaining tokens at the end that cannot form a complete block.

3.2 Pretrained Language Models

We select models to investigate how prior exposure to isiXhosa and isiZulu affects learning dynamics during CPT adaptation to these languages. Specifically, we select four encoder-only PLMs with different language coverage and levels of prior exposure to our target languages: English-only, massively multilingual, African-language, and Nguni-language. The models in these categories differ in the amount of isiXhosa and isiZulu text in their training corpora, which may influence how their learning dynamics develop during CPT. All selected models are open-source and publicly available. We use their largest available variants to improve comparability across the models, as described below.

English-Only Model. We use roberta-large [27] as our English-only model. RoBERTa was pretrained only on English, giving it no prior exposure to isiXhosa or isiZulu. This allows us to observe the learning dynamics of a model with no prior exposure to either target language.

Massively Multilingual Model. We select xlmr-large [9] as our massively multilingual model. XLM-R was pretrained on 100 languages, including isiXhosa but excluding isiZulu. This allows us to compare CPT learning dynamics for a seen and unseen target language within the same model.

African-Language Model. We use afriberta-large [36] as our African-language model. AfriBERTa was pretrained on 11 African languages, excluding English, isiXhosa, and isiZulu. This allows us to examine CPT learning dynamics for a model with broader African-language exposure but no prior exposure to either target language. AfriBERTa is also smaller and shallower than our other selected models, with 126 million parameters across 10 layers.

Nguni-Language Model. We select nguni-xlmr-large [32] as our Nguni-language model. Nguni-XLMR was adapted from XLM-R through CPT on isiXhosa, isiZulu, Siswati, and isiNdebele. This gives it direct prior exposure to both target languages, which allows us to examine how an already adapted model responds to additional CPT.

3.3 Continued Pretraining & Checkpoint Extraction

We continue pretraining each selected PLM monolingually on the preprocessed isiXhosa and isiZulu training splits from WURA. To

¹The code for our CPT procedure and learning dynamics analyses is available at <https://github.com/CodeWithSiya/cpt-learning-dynamics>.

do this, we adapt the source code² provided by Alabi et al. [2022] to continue pretraining each model on our target languages using the MLM objective [11]. We use a masking probability of 0.15 and follow the standard MLM masking scheme, where 80% of selected tokens are replaced with [MASK], 10% with a random token, and 10% remain unchanged [11]. We train the models using the AdamW optimiser [29] on NVIDIA L40S GPUs with 48 GB of VRAM and use a fixed random seed of 42 for reproducibility.

CPT Hyperparameters. For our CPT hyperparameters, we use the configuration from Meyer et al. [2024] as a starting point, as they demonstrate its effectiveness for adapting XLM-R to our target languages. Where necessary, we adjust this configuration when validation loss shows unstable training. We adopt their learning rate of 5×10^{-5} with no warmup. However, we found that AfriBERTa experienced a sharp spike in validation loss during the early stages of CPT under this configuration. Because prior work shows that training instability can occur in smaller models when using large learning rates [53], we tested a lower learning rate of 2×10^{-5} for AfriBERTa while keeping all other hyperparameters unchanged. This results in a more stable validation loss, so we adopt the lower learning rate for AfriBERTa.

We use an effective batch size of 80, following Meyer et al. [2024], as it was the largest batch size that fit within our available GPU memory across all models. We tested larger effective batch sizes of 92 and 128 to increase GPU utilisation, but neither was feasible across all models. We train all models for 200k steps. Given our small monolingual corpora, we found this training budget sufficient to observe meaningful learning dynamics during CPT adaptation to our target languages.

Checkpoint Extraction. We then extract model checkpoints at intervals adapted from Elhady et al. [2025]. During the first 10% of training steps, we save a checkpoint every 1% of steps to capture early behaviour changes in our PLMs. We use smaller checkpoint intervals here because model behaviour changes most rapidly during early training, and larger intervals risk missing important early trends [35]. For the remaining 90% of training, where changes occur more slowly, we save a checkpoint every 10% of steps. Including the initial model before CPT, this results in 20 checkpoints per CPT run and 160 checkpoints across all CPT runs.

Table 2: Downstream tasks and datasets used for isiXhosa and isiZulu.

Task	isiXhosa	isiZulu
NER	MasakhaNER 2.0	MasakhaNER 2.0
POS Tagging	MasakhaPOS	MasakhaPOS
NTC	MasakhaNEWS	ANTC

3.4 Fine-tuning on Downstream Tasks

Our first analysis examines how downstream task performance changes throughout CPT adaptation to our target languages. For each CPT checkpoint, we attach a randomly initialised task-specific

²<https://github.com/uds-lsv/afro-maft>

classification head and apply full fine-tuning on an NLU task to evaluate the checkpoint’s downstream performance. Following Devlin et al. [2018], for token classification tasks we apply a linear classification head to the final-layer representation of each input token. For sequence classification, we apply a similar classification head to the final-layer representation of the [CLS] token, which represents the entire input sequence. We evaluate each CPT checkpoint on NER, POS tagging, and NTC, as summarised in Table 2 and discussed below.

Tasks & Datasets. NER is a token-level sequence labelling task that identifies spans of text corresponding to named entities such as people and locations [55]. We fine-tune our checkpoints on MasakhaNER 2.0 [3], a human-annotated NER dataset covering 20 African languages, including isiXhosa and isiZulu. POS tagging is another token-level labelling task that assigns a grammatical category, such as noun or adjective, to each token in a sequence [12]. We use MasakhaPOS [12], a dataset covering 20 diverse African languages, including our target languages. Our final task is NTC, a text classification task that categorises news articles into categories such as sport and politics [2]. Since isiXhosa and isiZulu are covered by different NTC datasets, we use MasakhaNEWS [2] and ANTC [4] for isiXhosa and isiZulu, respectively. To make learning dynamics easier to compare across tasks and models, we report macro-F1 for all three tasks, following Müller-Eberstein et al. [2023].

Fine-Tuning Hyperparameters. As with CPT, we use the fine-tuning hyperparameters reported by Meyer et al. [2024] as our starting point and adjust them based on fine-tuning validation performance. We found that this configuration caused fine-tuning instability on some CPT checkpoints. Because Mosbach et al. [2020] show that high learning rates and the absence of warmup can cause fine-tuning instability for BERT-style models, we reduced our initial learning rate to 2.0×10^{-5} and added a linear warmup over the first 10% of training steps to improve training stability. This stabilised fine-tuning across the affected checkpoints. We fine-tune each checkpoint for five epochs, as longer fine-tuning caused many checkpoints to overfit, with training loss continuing to improve while validation loss stopped improving. Within our five-epoch budget, we select the epoch with the best validation performance for each checkpoint and report its corresponding test-set performance. We apply the same fine-tuning hyperparameters across all three downstream tasks and repeat fine-tuning for each checkpoint using five random seeds (see Table A2). We plot the mean macro-F1 across the five fine-tuning seeds.

3.5 Cross-Lingual Geometry Analysis

Our second analysis investigates how the geometry of our models’ cross-lingual representations changes during CPT. We examine two geometric properties to help explain these representational changes: cross-lingual alignment and isotropy. Because greater alignment between representations of our target languages and English may support cross-lingual transfer, we use cosine similarity to assess whether isiXhosa and isiZulu become more closely aligned with English during CPT. We then use isotropy to assess how effectively the models use their available representation space.

Dataset. For our geometric analyses, we use parallel isiXhosa-English and isiZulu-English sentence pairs from FLORES-200 [50], because it has been used in prior work on representation analysis [24, 31]. FLORES-200 is a high-quality multilingual evaluation corpus containing parallel translations across over 200 languages, including our target languages and English. Its parallel structure allows us to compare representations of semantically equivalent sentences in English and our target languages. We use the devtest split as a held-out set distinct from WURA, which provides 1,012 aligned sentence pairs for each target language and English.

Methods. We compute geometric metrics at every layer of our selected PLMs because prior work shows that representations vary across layers [15, 31], which allows us to identify where cross-lingual geometry changes occur within the models. We also track how these representations develop over time by analysing five checkpoints spanning the early, middle, and late stages of CPT. Specifically, we analyse checkpoints at steps 0, 40k, 100k, 140k, and 200k. For each checkpoint, we extract the token-level representations of every FLORES-200 sentence at each layer using the model’s standard tokeniser. Following Conneau et al. [2019], we obtain a sentence-level representation by averaging the token representations at each layer, which we denote by $s^{(l)}$ for layer l . We exclude special tokens from this average and use these sentence representations to compute cosine similarity and isotropy.

Cosine Similarity. We measure cross-lingual alignment using cosine similarity between each target-language sentence representation and its parallel English representation. For layer l , we compute the average cosine similarity across all N parallel sentence pairs as

$$\text{Similarity}(l) = \frac{1}{N} \sum_{(i,j)} \frac{s_i^{(l)} \cdot s_j^{(l)}}{\|s_i^{(l)}\| \|s_j^{(l)}\|},$$

where $s_i^{(l)}$ and $s_j^{(l)}$ are the layer- l representations of a target-language sentence i and its English translation j , respectively, and N is the number of parallel sentence pairs. Higher cosine similarity indicates greater alignment between semantically equivalent sentences in the two languages.

Isotropy. Cosine similarity measures alignment between individual sentence pairs, but it does not describe how uniformly the joint representation space is used. Following Rudman et al. [2022], we measure isotropy using IsoScore, which quantifies how evenly representations use the available dimensions of this space. For each layer, we combine the English and target-language sentence representations into a single set X and compute IsoScore over this joint representation space as

$$\text{IsoScore}(l) = \frac{(d - \delta(X)^2(d - \sqrt{d}))^2 - d}{d(d - 1)},$$

where d is the representation dimension and $\delta(X) \in [0, 1]$ is the isotropy defect, which measures how far the variance distribution of X deviates from uniformity. IsoScore ranges from 0 to 1, with higher values indicating that the English and target-language

representations use the available representation space more uniformly. A low IsoScore alongside high cosine similarity suggests that the observed similarity may partly reflect anisotropy rather than meaningful cross-lingual alignment.

Because our work focuses on interpretability and analysing internal model representations, we do not perform statistical significance testing or conventional error analysis. Instead, we strengthen and validate our conclusions by determining whether the observed learning patterns are consistent across multiple models, tasks, and languages. We discuss the ethical considerations of our work in Section 6.

4 DOWNSTREAM PERFORMANCE ANALYSIS

We now analyse how downstream task performance changes during CPT adaptation to isiXhosa and isiZulu. We examine these changes at two levels. First, we analyse how each model’s macro-F1 changes across CPT training steps for each downstream task (see Figure 2). We then analyse performance on individual classes within these tasks, such as specific POS tags and news topics, to examine how different categories respond to CPT (see Figure 3).

4.1 Downstream Performance During CPT

Figure 2 shows that downstream performance generally follows three stages during CPT. Most improvements occur early in training, followed by slower gains, with performance becoming largely stable in the later stages of CPT. However, the magnitude of these changes differs across models and tasks, with models that have less prior exposure to isiXhosa and isiZulu generally showing larger improvements during CPT.

Stage 1: Rapid Early Adaptation. We observe that the largest downstream performance changes generally occur within the first 40,000 CPT steps. This is most visible in NTC performance, where RoBERTa and AfriBERTa show clear early improvements for isiXhosa. XLM-R also improves for both languages, although the increase is larger for isiZulu. This pattern is consistent with prior work on CPT learning dynamics. Elhady et al. [2025] identify a critical period at the beginning of language-adaptive CPT, during which the shift to a new language produces the largest changes in model behaviour and parameters. Although their study examines decoder-only models and different downstream abilities, our results suggest a similar pattern for encoder-only models, with the largest downstream changes occurring early in CPT.

The amount of early change also appears to depend on prior language exposure. XLM-R shows limited change for isiXhosa, which was included in its original pretraining data. In contrast, it shows noticeably larger early improvements for isiZulu, especially in NTC performance. Since isiZulu was absent from its pretraining data, this contrast suggests that direct prior exposure to a target language may reduce the additional adaptation required during CPT. Nguni-XLMR provides a stronger example of this pattern. Having already been continually pretrained on isiXhosa and isiZulu, it starts with strong downstream performance and shows little further improvement for isiXhosa. Our findings are consistent with Elhady et al. [2025], who find that models with stronger initial target-language abilities are less affected by the transition to CPT,

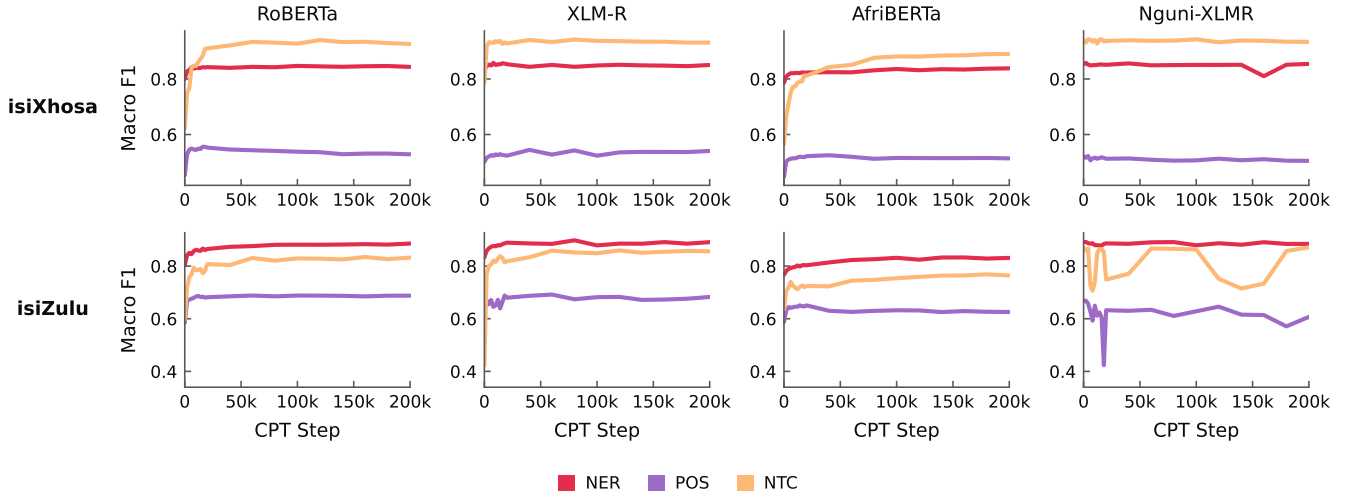


Figure 2: Downstream macro-F1 across CPT checkpoints for NER, POS tagging, and NTC in isiXhosa and isiZulu for RoBERTa, XLM-R, AfriBERTa, and Nguni-XLMR.

and with Meyer et al. [2024], who found larger gains when adapting XLM-R to Nguni languages absent from its original pretraining data.

The size of the early changes also differs by task. NTC generally shows the largest improvements, while POS tagging changes comparatively little and NER typically shows smaller gains. Since NTC is domain-specific, models may begin with lower performance before improving as they adapt to news-specific language during CPT. In addition, because WURA consists primarily of news text, CPT may adapt models to both the target language and the news domain before fine-tuning. Gururangan et al. [2020] show that adaptation to in-domain text can improve downstream performance in both low- and high-resource settings, which may explain why NTC benefits more than NER and POS tagging.

Stage 2: Slower Improvement. Changes in downstream performance slow down after the initial period of rapid adaptation. We observe this slower improvement from CPT steps 40,000 to 100,000. NTC performance continues to improve for RoBERTa and AfriBERTa in both languages and for XLM-R in isiZulu, but at a slower rate than during early CPT. In contrast, NER and POS tagging show smaller changes during this stage, with performance remaining largely stable across most models and both languages. Nguni-XLMR also remains comparatively stable for isiXhosa during this stage. This transition from rapid early change to slower subsequent adaptation is consistent with Elhady et al. [2025], who find that the largest parameter shifts and changes in existing capabilities occur during the initial stages of CPT, followed by slower changes for the remainder of training. This pattern also reflects pretraining learning dynamics, as Müller-Eberstein et al. [2023] find a similar rapid early learning phase followed by slower gains when pretraining encoder-only models from scratch.

However, across CPT, Nguni-XLMR behaves differently for isiZulu. NTC is highly unstable, with several large drops followed by recovery. POS tagging also shows a sharp early drop followed by smaller fluctuations, while NER remains largely stable throughout.

This suggests that additional CPT may interfere with knowledge already present in the model, although the instability appears task-specific rather than a general loss of model ability. The sharp initial declines in NTC and POS tagging may reflect catastrophic forgetting, while the larger fluctuations in NTC may indicate that this domain-specific task is more strongly affected by additional CPT. This early degradation is consistent with Elhady et al. [2025], who observe catastrophic forgetting during the critical period of language-adaptive CPT when English is not included. Since our CPT similarly uses only isiZulu text, the sharp early declines in Nguni-XLMR may reflect a similar effect. Notably, this instability is absent in the other three models, which lack comparable prior exposure to isiZulu.

Stage 3: Stabilisation. During the final stages of CPT, most models largely stabilise. However, some models continue to improve slightly, such as AfriBERTa on NTC in both languages. We observe this general stabilisation after 100,000 CPT steps. RoBERTa, XLM-R, and AfriBERTa generally remain close to the performance reached earlier in training across both isiXhosa and isiZulu, while Nguni-XLMR is similarly stable for isiXhosa. Nguni-XLMR shows a temporary dip in isiXhosa NER performance before recovering, which is consistent with our earlier finding that additional CPT can disrupt performance in an already adapted model. Since most downstream gains occur earlier in training, our results suggest that substantially fewer CPT steps may be sufficient for encoder-only models to reach most of their eventual downstream performance, indicating diminishing returns from extended CPT once performance has stabilised. This aligns with findings from pretraining from scratch, where Chiang et al. [2020] show that encoder-only models can reach near-final downstream performance after only a fraction of their full pretraining.

4.2 Class-Level Performance During CPT

In this section, we analyse class-level performance within our downstream tasks to better understand the aggregate trends observed

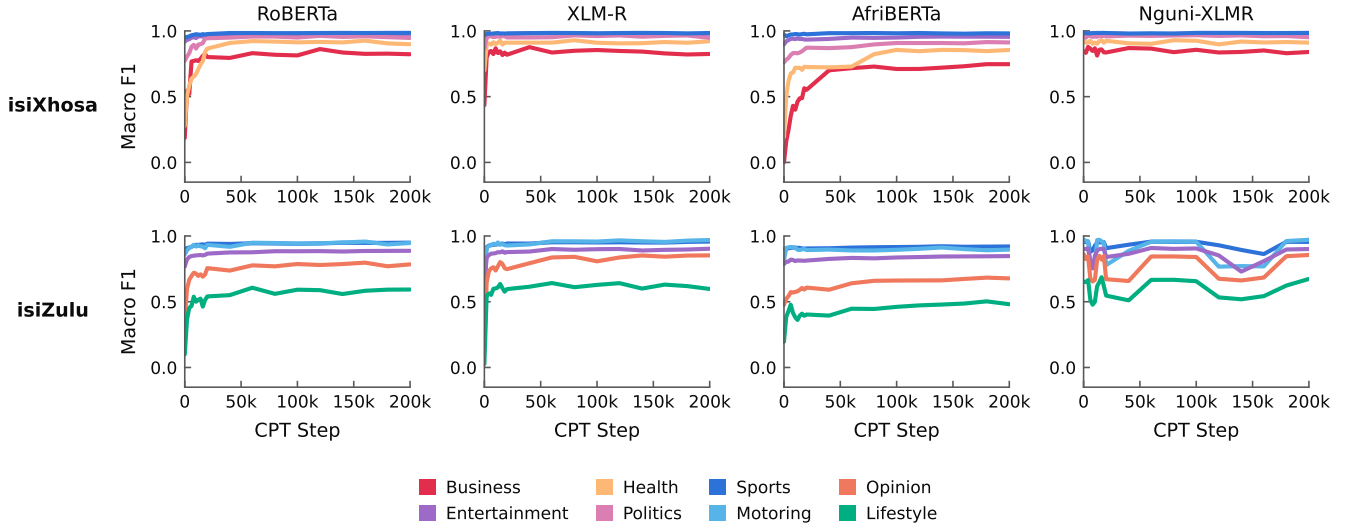


Figure 3: NTC macro-F1 across CPT checkpoints for individual news categories in isiXhosa and isiZulu for RoBERTa, XLM-R, AfriBERTa, and Nguni-XLMR.

in Section 4.1. Following Müller-Eberstein et al. [2023], we use class-level performance to identify changes that aggregate metrics may hide, which allows us to see which types of linguistic knowledge continue to change during CPT. Figure 3 presents the NTC results, while Appendix C provides the corresponding NER and POS tagging results.

News Topic Classification. Figure 3 shows that class-level NTC performance broadly follows the overall macro-F1 trends, with most improvements occurring early in CPT before slowing and largely stabilising. Nguni-XLMR on isiZulu again differs from this pattern, with several categories decreasing and recovering during CPT. This shows that Nguni-XLMR’s overall NTC instability affects multiple news categories and is not limited to a single class. The class-level results also show differences in how individual categories respond to CPT. Categories that start with relatively strong performance, such as Sports and Entertainment, generally show limited improvement across both languages. In contrast, categories that start with lower performance, such as Business and Health in isiXhosa, often improve more. This suggests that categories models initially find difficult account for more of the overall NTC improvement observed in Figure 2.

One possible explanation is that some categories can be identified from distinctive keywords. Sports, for example, may contain frequent references to particular sports, teams, or competitions, whereas Business covers a broader range of topics and may require more domain-specific vocabulary and contextual information. CPT may therefore provide greater benefit for categories that depend on broader target-language and domain-specific knowledge.

Named Entity Recognition. NER broadly follows the overall downstream patterns in Figure 2, although the class-level changes are smaller than those observed for NTC (see Figure C1). Most entity classes improve early in CPT and then remain relatively stable. However, RoBERTa and AfriBERTa continue to show noticeable

improvements for several isiZulu entities throughout CPT. This is consistent with the broader pattern that models with less prior exposure to the target language continue to benefit from additional CPT. Nguni-XLMR on isiXhosa also shows a temporary decrease across several entity classes during later CPT, which corresponds to the dip observed in its overall NER performance.

The relative ordering of entity classes is also consistent across models and languages. PER generally achieves the strongest performance, while DATE and ORG tend to perform worse. One possible explanation is that person names often have clearer cues that help models recognise them as names, while organisations can have more varied names containing common words. Dates can also be expressed in many ways, which may make them harder to recognise consistently. The continued improvements in some ORG classes may therefore come from encountering more organisation names and contexts during CPT. This aligns with Müller-Eberstein et al. [2023], who find that open-domain entity classes, such as organisations, can continue improving later in pretraining from scratch as models encounter new entities.

POS Tagging. For readability, we group the POS tags, which follow the Universal POS tagset [40], into the categories shown in Table B1 in Appendix B. POS tagging generally shows smaller class-level changes during CPT than NTC and NER (see Figure C2). Most grammatical categories improve slightly at the beginning of CPT and then remain close to their initial performance, with little change in their relative ordering across models and languages. This aligns with the stable overall POS performance observed in Section 4.1. One exception is the Modifiers category, which improves early before quickly stabilising for RoBERTa and AfriBERTa in both languages. Modifiers such as adjectives and adverbs often depend on surrounding words and grammatical context, so additional target-language exposure may help models recognise these patterns more accurately. This is consistent with Chiang et al. [2020], who show

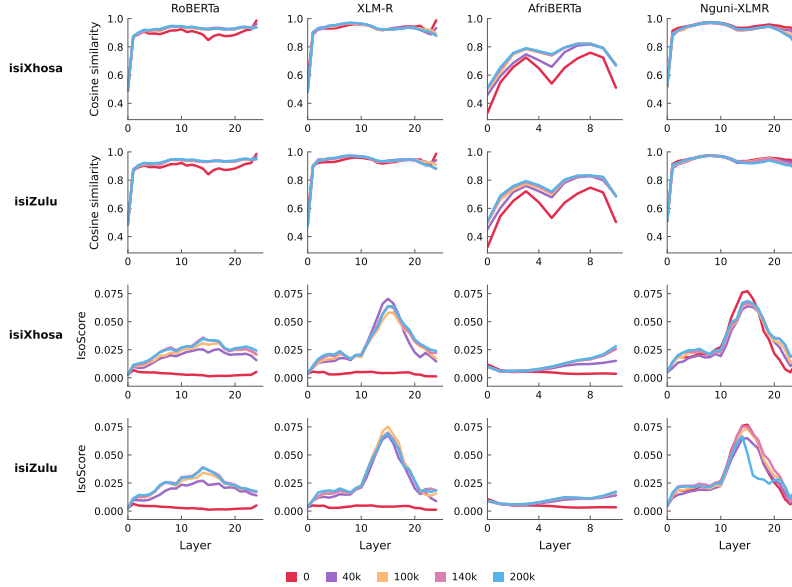


Figure 4: Layer-wise cross-lingual alignment and isotropy across selected CPT checkpoints for isiXhosa and isiZulu for RoBERTa, XLM-R, AfriBERTa, and Nguni-XLMR.

that different POS classes are learned at different rates during pre-training, with function words learned earlier than modifiers such as adjectives and adverbs. Punctuation is also less stable for some isiZulu models, while Nguni-XLMR again shows fluctuations across several categories, consistent with its overall POS instability.

The stability of POS performance suggests that CPT provides relatively little additional benefit for many grammatical categories in isiXhosa and isiZulu. This is notable even for models with little or no direct exposure to the target languages before CPT, which suggests some transfer of grammatical knowledge across languages. Our findings suggest that CPT provides less additional benefit for grammatical knowledge than for the language- and domain-specific knowledge required for tasks such as NTC.

5 GEOMETRIC ANALYSIS

Our second analysis examines the cross-lingual representations of our selected PLMs during CPT adaptation to isiXhosa and isiZulu. Figure 4 shows how cross-lingual alignment and isotropy change across model layers at our five selected CPT checkpoints. First, we analyse how these properties vary across model layers. We then analyse how they change across CPT checkpoints to provide an internal perspective that complements our downstream analysis.

5.1 Cross-Lingual Geometry Across Layers

We first examine how cross-lingual alignment and isotropy change across the layers of our selected PLMs at different CPT checkpoints. RoBERTa, XLM-R, and Nguni-XLMR show similar cosine similarity patterns across both languages. Alignment increases sharply within the first few layers and remains high across most intermediate and upper layers. XLM-R and Nguni-XLMR then show a clearer decrease towards the final layers, while RoBERTa remains relatively stable. The intermediate-layer peaks observed for XLM-R

and Nguni-XLMR are consistent with Kargaran et al. [2025], who find strong English–non-English alignment in the intermediate layers of decoder-only models and show that this is associated with improved cross-lingual transfer and multilingual downstream performance. Our results suggest that a similar layer-wise organisation also appears in encoder-only models during CPT.

Isotropy shows a related pattern. RoBERTa, XLM-R, and Nguni-XLMR again show similar trends across both languages, although the changes are less pronounced for RoBERTa. Isotropy generally increases through the middle or upper-middle layers before decreasing towards the final layer. This suggests that the models make the most effective use of their representational capacity in the middle and upper-middle hidden layers before representational degeneration becomes more pronounced towards the final layers. Meyer and Buys [2026] similarly find that representational degeneration in low-resource languages is most consistent in the final layers of decoder-only LLMs. Although our analysis uses joint English–target-language sentence representations rather than within-language token geometry, both results show that representational degeneration becomes more pronounced towards the final layers.

AfriBERTa follows a different pattern. Both its cosine similarity and IsoScore change more gradually across its smaller number of layers, and its isotropy does not show the same pronounced middle-layer peak seen in XLM-R and Nguni-XLMR. Its lower cross-lingual alignment may reflect its pretraining data, which excludes English and our target languages. However, AfriBERTa is also smaller and has fewer hidden layers than the other models, which may contribute to its different layer-wise pattern. This suggests that differences in layer-wise geometry may reflect broader pretraining language coverage and model architecture.

5.2 Cross-Lingual Geometry During CPT

Our final analysis examines how cross-lingual alignment and isotropy change across CPT checkpoints. We organise our analysis into two phases to compare how downstream performance and cross-lingual geometry develop during CPT.

Geometry During Early CPT. We observe the largest changes in cross-lingual geometry between the initial model and the first analysed CPT checkpoint at step 40,000. This is most visible in IsoScore. RoBERTa and XLM-R begin with highly anisotropic joint English–target-language representation spaces, but IsoScore increases after CPT begins for both languages. In contrast, Nguni-XLMR begins with a higher IsoScore, which suggests that its earlier Nguni-language CPT had already produced a more isotropic representation space. However, its IsoScore decreases after an additional 40,000 CPT steps, indicating increased representational degeneration. This further suggests that additional CPT on an already adapted model may disrupt previously established geometric structure, which may help explain the limited and unstable downstream changes observed for Nguni-XLMR. These early geometric changes also reflect our downstream results, where the largest performance changes occur early in CPT, and are consistent with the critical early period reported by Elhady et al. [2025].

Cosine similarity shows a complementary pattern. RoBERTa and XLM-R already show high English–target-language similarity across most hidden layers at step 0, although this occurs alongside low IsoScore, which suggests that some of the similarity may reflect anisotropy. Gao et al. [2019] similarly show that high cosine similarity can result from representational degeneration. By step 40,000, cosine similarity increases most clearly for RoBERTa and AfriBERTa, providing stronger evidence of increased cross-lingual alignment. AfriBERTa’s lower initial similarity may partly reflect its pretraining data, which includes neither English nor our target languages. In contrast, Nguni-XLMR shows a slight decrease in cosine similarity, which suggests that additional CPT may initially reduce cross-lingual alignment in an already adapted model. Overall, these findings are consistent with our downstream finding that models with less prior target-language exposure tend to show larger changes during CPT.

Geometry During Later CPT. After 40,000 steps, increases in isotropy generally become much smaller for RoBERTa, XLM-R, and AfriBERTa. Their later IsoScore curves largely overlap, which suggests that most improvements in how effectively the representation space is used occur early in CPT. Nguni-XLMR on isiXhosa follows the same general pattern, but its IsoScore remains below its initial level following the early decrease, suggesting that this representational degeneration persists through later checkpoints. For isiZulu, IsoScore partially recovers at the 100,000 and 140,000 checkpoints before decreasing again at step 200,000. This partly reflects the decreases and recoveries observed in downstream performance, although the two patterns do not fully align. This is consistent with prior work showing that the relationship between representational geometry and downstream performance is not always clear [31].

Cosine similarity also becomes relatively stable after 40,000 steps, with later checkpoint curves increasing slightly and largely overlapping across models and languages. For Nguni-XLMR, however, this stable level remains below its initial alignment following the early

decrease. Overall, our results suggest that the largest changes in cross-lingual geometry occur during early CPT. CPT can increase alignment between the target languages and English while also reducing anisotropy in the joint representation space, although the magnitude and layer-wise location of these changes vary across models.

6 CONCLUSIONS

In this study, we investigated the learning dynamics of four encoder-only PLMs during CPT adaptation to isiXhosa and isiZulu using downstream evaluation and cross-lingual representation analysis. We found that most downstream improvements occur early in CPT before slowing and largely stabilising, with NTC generally benefiting more than NER and POS tagging. Models with less prior target-language exposure generally show larger downstream and geometric changes, while models with stronger prior exposure change less and may become unstable under additional CPT, as observed for Nguni-XLMR. Our geometric analysis further shows that CPT can improve English–target-language alignment and isotropy, with both generally strongest in intermediate and upper-middle layers before representational degeneration becomes more pronounced towards the final layers.

Our findings extend learning-dynamics research beyond pre-training from scratch to CPT-based adaptation of encoder-only models for low-resource African languages. They also show that CPT does not affect all models, tasks, or hidden layers equally. Since most downstream gains occur early, shorter CPT budgets may achieve similar downstream performance, which can reduce the computational costs associated with adapting models to new languages. This is valuable in low-resource settings, where compute is often limited.

Our study is limited to two closely related languages, four encoder-only models, and geometric analysis using a single parallel evaluation corpus. Future work could examine additional low-resource languages and model architectures to determine how widely these dynamics generalise. Another promising direction is to apply the principle of actionable interpretability by using learning-dynamics findings to improve CPT itself, for example through model-specific stopping criteria or interventions targeted at stages or layers where useful adaptation occurs.

ETHICS STATEMENT

This study presents no significant ethical risks. We use encoder-only models designed for language understanding tasks such as text classification and linguistic annotation, which limits their potential for harmful use. The datasets used may reflect biases in their source material, but we do not publicly release the CPT checkpoints produced in this study.

ACKNOWLEDGEMENTS

We thank Dr Francois Meyer for his continued guidance and support throughout this research. Computations were performed using facilities provided by the University of Cape Town’s ICTS High Performance Computing team (<https://hpc.uct.ac.za>). This work is based on research supported by the National Research Foundation of South Africa (Reference: PMDS250930369454).

AI USAGE STATEMENT

Generative AI was used to provide feedback on writing, including clarity, structure, and organisation. We used Grammarly for proofreading and grammar suggestions, which we reviewed before applying. We used ChatGPT for general feedback on writing quality and structure, as well as assistance with L^AT_EX formatting. No AI-generated text was copied directly into the document. All content was written and revised by the author, with AI used only as a supporting tool.

REFERENCES

- [1] David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajudeen 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 Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthulu. 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 Marnette, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3053–3070. doi:10.18653/v1/2022.naacl-main.223
- [2] David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Oduwayo Ogundepo, Bonaventure F.P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Sabah Al-Azzawi, Blessing K. Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde O. Ajayi, Tatiana Moteu Ngoli, Brian Odhiambo, Abraham Toluwase Owoadunni, Nnaemeka C. Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed S. Abdullahi, Mesay Gemedi Yigezu, Tajudeen Gwadabe, Idris Abdulmumin, Mahlet Taye Bame, Oluwabusayo O. Awoyomi, Iyanuoluwa Shode, Tolulope Anu Adelani, Habiba Abdulganiny Kailani, Abdul Hakeem Omotayo, Adetola Adeeko, Afolabi Abebe, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Raphael Ogbu, Chinedu E. Mbonu, Chiamaka I. Chukwunke, Samuel Fanijo, Jessica Ojo, Oyinkansola F. Awosan, Tadesse Kebede Guge, Sakayo Toadoun Sari, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyah Oduwole, Kanda P. Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos G. Nugussie, Abdulmejid Tuni Johar, Shafie Abdi Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Twagirayezu, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News Topic Classification for African languages. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Long Papers, IJCNLP-AACL 2023 1* (2023), 144–159. doi:10.18653/V1/2023.IJCNLP-MAIN.10
- [3] David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F.P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire M. Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajudeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwunke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Oduwayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022* (2022), 4488–4508. doi:10.18653/V1/2022.EMNLP-MAIN.298
- [4] 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/>
- [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] 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>
- [8] 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>
- [9] 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>
- [10] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging Cross-lingual Structure in Pretrained Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (11 2019), 6022–6034. doi:10.18653/v1/2020.acl-main.536
- [11] 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>
- [12] Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba O. Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F.P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajudeen Gwadabe, Elvis Tchiaye Mboning, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Anu Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudza Gotsa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African languages. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1 (2023), 10883–10900. doi:10.18653/V1/2023.ACL-LONG.609
- [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] 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>
- [15] Kavin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (2019), 55–65. doi:10.18653/V1/D19-1006
- [16] 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
- [17] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie Yan Liu. 2019. Representation Degeneration Problem in Training Natural Language Generation Models. *7th International Conference on Learning Representations, ICLR 2019* (7 2019). <https://arxiv.org/pdf/1907.12009>
- [18] Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024. Anisotropy Is Inherent to Self-Attention in Transformers. *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference* 1 (2024), 35–48. doi:10.18653/V1/2024.EACL-LONG.3
- [19] 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
- [20] Danny Hernandez, Tom Brown, Tom Conerly, Nova Dassarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish Anthropic. 2022. Scaling Laws and Interpretability of Learning from Repeated Data. (5 2022). <https://arxiv.org/pdf/2205.10487>
- [21] 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>
- [22] Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring Anisotropy and Outliers in Multilingual Language Models for Cross-Lingual Semantic Sentence Similarity. (6 2023). <https://arxiv.org/pdf/2306.00458>
- [23] Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding Cross-Lingual Alignment – A Survey. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (4 2024), 10922–10943. doi:10.18653/v1/2024.findings-acl.649
- [24] Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2025. MEXA: Multilingual Evaluation of English-Centric LLMs via Cross-Lingual Alignment. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2025), 27001–27023. doi:10.18653/v1/2025.FINDINGS-ACL.1385
- [25] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmongkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroko Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidek Mirzakhlov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhiile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics* 10 (2 2022), 50–72. doi:10.1162/tacl_a_00447
- [26] 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>
- [27] 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>
- [28] Anri Lombard, Simbarashe Mawere, Temi Aina, Ethan Wolff, Sbonelo Gumede, Elan Novick, Francois Meyer, and Jan Buys. 2026. MzansiText and MzansiLM: An Open Corpus and Decoder-Only Language Model for South African Languages. (3 2026). <https://arxiv.org/pdf/2603.20732>
- [29] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. *7th International Conference on Learning Representations, ICLR 2019* (11 2017). <https://arxiv.org/pdf/1711.05101>
- [30] Francois Meyer and Jan Buys. 2025. The Learning Dynamics of Subword Segmentation for Morphologically Diverse Languages. (11 2025). <http://arxiv.org/abs/2511.09197>
- [31] Francois Meyer and Jan Buys. 2026. The Geometry of Low-Resource Language Representations. (8 2026). <https://arxiv.org/pdf/2608.23358>
- [32] 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)*, 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/>
- [33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* (1 2013). <https://arxiv.org/pdf/1301.3781>
- [34] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *ICLR 2021 - 9th International Conference on Learning Representations* (6 2020). <https://arxiv.org/pdf/2006.04884>
- [35] 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>
- [36] Kelechi Ogueji, Yuxin Zhu, Jimmy Lin, and David R Cheriton. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resource Languages. (2021), 116–126. <https://github.com/google-research/bert/>
- [37] 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>
- [38] Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaghene 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
- [39] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (2018), 2227–2237. doi:10.18653/V1/N18-1202
- [40] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A Universal Part-of-Speech Tagset. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012* (4 2011), 2089–2096. doi:10.6337/3pjeez8kmmkhz
- [41] 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>
- [42] Laurette Pretorius and Sonja Bosch. 2009. Exploiting Cross-Linguistic Similarities in Zulu and Xhosa Computational Morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages*, Lori Levin, John Kiango, Judith Klavans, Guy De Pauw, Gilles-Maurice de Schryver, and Peter Waiganjo Wagacha (Eds.). Association for Computational Linguistics, Athens, Greece, 96–103. <https://aclanthology.org/W09-0714>
- [43] 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
- [44] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassiere, Richard Powell, George Van Den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien De Masson D'autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego De Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. (12 2021). <https://arxiv.org/pdf/2112.11446>
- [45] Sara Rajae and Mohammad Taher Pilehvar. 2021. A Cluster-based Approach for Improving Isotropy in Contextual Embedding Space. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 2 (2021), 575–584. doi:10.18653/V1/2021.ACL-SHORT.73
- [46] William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the Uniformity of Embedding Space Utilization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (4 2022), 3325–3339. doi:10.18653/v1/2022.findings-acl.262
- [47] 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
- [48] Statistics South Africa. 2024. *Census 2022 in brief*. Technical Report. Stats SA. Table 3.8 & Figure 3.9. <https://www.statssa.gov.za/publications/Census2022inBrief/Census2022inBriefJune2024.pdf>
- [49] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi

- Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiayi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaoxue Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyang Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoquan Mao, Yuan Mei, Xin Men, Mingqi Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yungeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenchi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjie Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuyue Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiawei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijiang Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanlei Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiye Zhuang, and Xinxing Zu. 2026. Kimi K3: Open Frontier Intelligence. (7 2026). <https://arxiv.org/pdf/2607.24653>
- [50] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. (7 2022). <https://arxiv.org/pdf/2207.04672>
- [51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal 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, Madsen Khabsa, Isabel Kloumann, Artem Korneev, 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>
- [52] 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>
- [53] Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. 2023. Small-scale proxies for large-scale Transformer training instabilities. *12th International Conference on Learning Representations, ICLR 2024* (9 2023). <https://arxiv.org/pdf/2309.14322>
- [54] 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>
- [55] Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. 2145-2158 pages. <https://aclanthology.org/C18-1182/>
- [56] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuxiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. (5 2025). <https://arxiv.org/pdf/2505.09388>
- [57] 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>

A HYPERPARAMETERS

Table A1: CPT hyperparameter configurations used for our selected PLMs. The same configuration is used across models except for the lower learning rate used for AfriBERTa to improve training stability.

Hyperparameter	RoBERTa	XLM-R	AfriBERTa	Nguni-XLMR
Learning rate	5.0×10^{-5}	5.0×10^{-5}	2.0×10^{-5}	5.0×10^{-5}
Optimiser	AdamW	AdamW	AdamW	AdamW
Batch size (per device)	10	10	10	10
Gradient accumulation steps	8	8	8	8
Effective batch size	80	80	80	80
Total training steps	200,000	200,000	200,000	200,000
Warmup steps	0	0	0	0
Maximum sequence length	512	512	512	512
Random seed	42	42	42	42

Table A2: Downstream fine-tuning hyperparameter configurations used for NER, POS tagging, and NTC. We use the same configuration across all three downstream tasks.

Hyperparameter	NER	POS	NTC
Learning rate	2.0×10^{-5}	2.0×10^{-5}	2.0×10^{-5}
Optimiser	AdamW	AdamW	AdamW
Epochs	5	5	5
Batch size	32	32	32
Warmup ratio	0.1	0.1	0.1
Random seeds	{42, 123, 456, 789, 1738}		

B PART-OF-SPEECH TAGGING GROUPS

Table B1: Grouping of Universal Dependencies POS tags used in our per-class POS tagging analysis. We combine semantically and syntactically related tags into broader categories to reduce the number of individual classes and make changes in POS tagging downstream performance across CPT checkpoints easier to interpret.

Group	Universal Dependencies Tags
Nouns	NOUN, PROPN
Verbs	VERB, AUX
Modifiers	ADJ, ADV
Function	ADP, DET, PRON, CCONJ, SCONJ, PART
Punctuation	PUNCT, SYM
Other	INTJ, NUM, X

C DOWNSTREAM PERFORMANCE ANALYSIS RESULTS

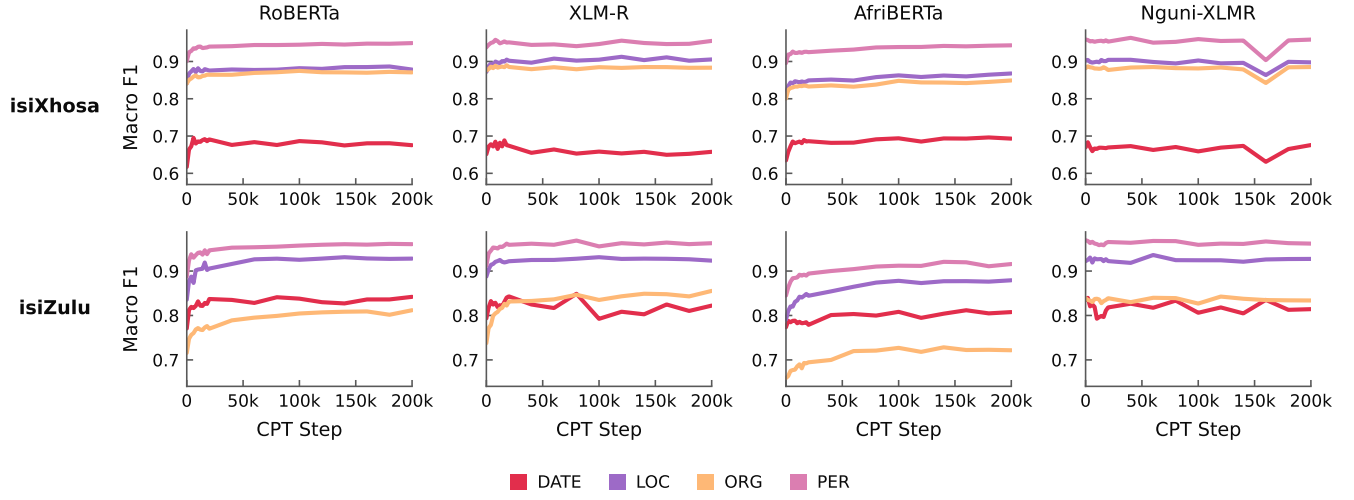


Figure C1: NER macro-F1 across CPT checkpoints for individual named entities in isiXhosa and isiZulu for RoBERTa, XLM-R, AfriBERTa, and Nguni-XLMR.

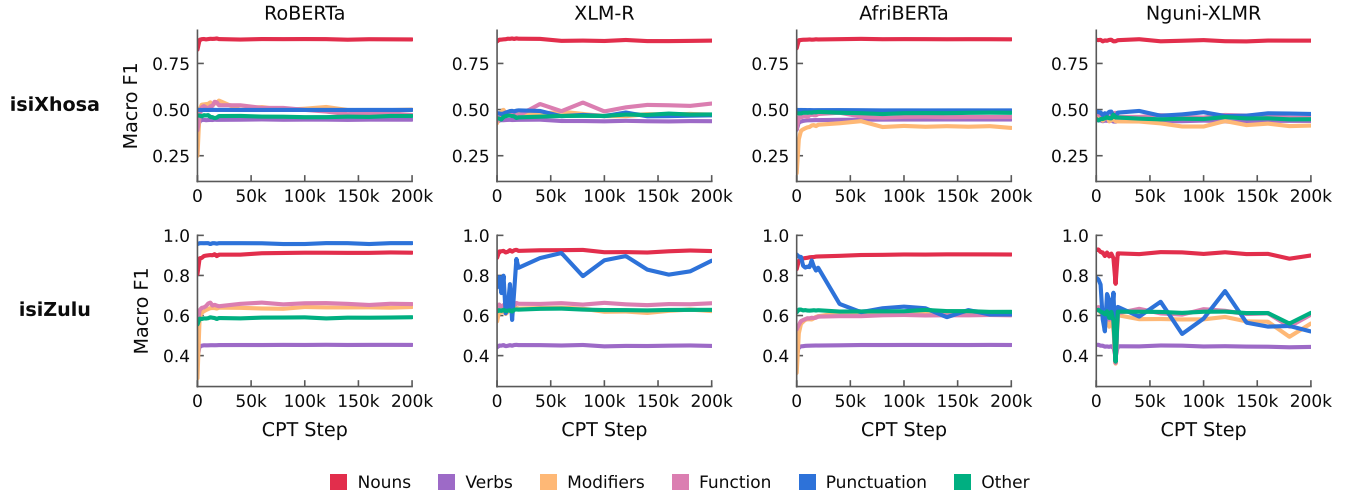


Figure C2: POS tagging macro-F1 across CPT checkpoints for grouped grammatical categories in isiXhosa and isiZulu for RoBERTa, XLM-R, AfriBERTa, and Nguni-XLMR.

D CONTINUED PRETRAINING LOSS CURVES

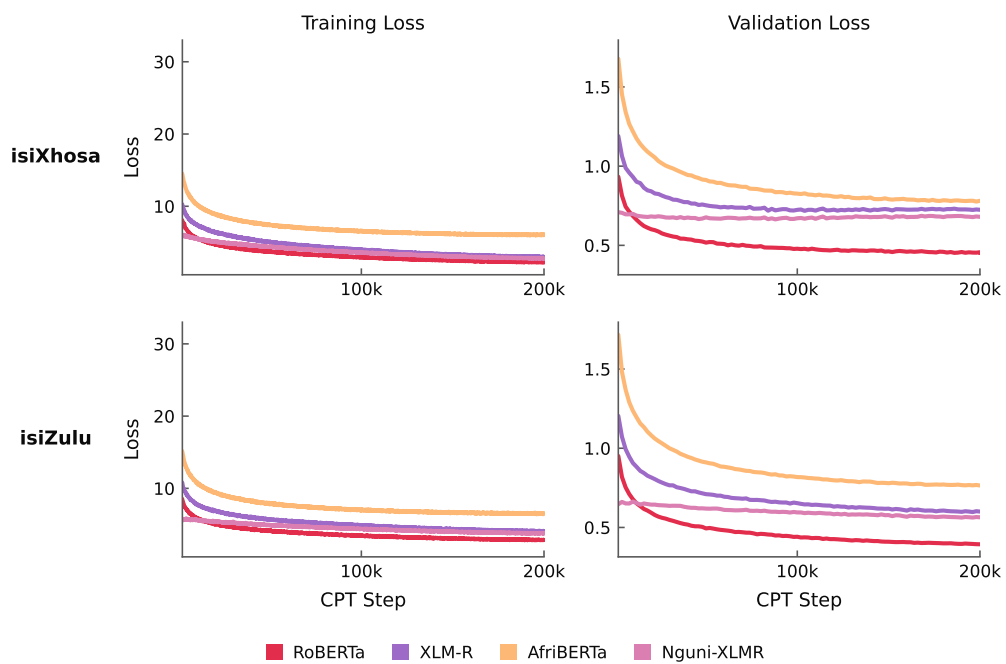


Figure D1: CPT training and validation loss across training steps for RoBERTa, XLM-R, AfriBERTa, and Nguni-XLMR on isiXhosa and isiZulu.