JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

Parameter-Efficient Fine-Tuning for Equitable Named Entity Recognition in African Languages: A Lora-Based Approach

[*1]Erewa Omasan, [2]Oluwatobi Noah Akande, [3]Yusuf Salisu Ibrahim, [4]Joshua Abah, [5]Julius Makinde

[1,2,3,4] Computer Science Department, Nile University of Nigeria, Abuja, Nigeria
[5]Computer Science Department, Baze University of Nigeria.
omaserewa@gmail.com

ABSTRACT

Named Entity Recognition (NER) remains a challenge for African languages, where limited annotated data, typological diversity, and the computational cost of full fine-tuning restrict practical deployment. This research investigates whether LoRA combined with balanced multilingual sampling can simultaneously address computational efficiency and cross-language equity for African language NER. Using the MasakhaNER 2.0 dataset, we fine-tuned the pre-trained XLM-RoBERTa model using LoRA adapters with a rank of 32 covering Hausa, Yoruba, Swahili, Igbo, Kinyarwanda, and Nigerian Pidgin. A balanced sampling strategy equalises training-set contributions across languages at each gradient step. Statistical reliability is assessed through joint multi-seed evaluation across three random seeds, and sampling-strategy effects are isolated in a controlled balanced versus unbalanced ablation. The LoRA model surpasses the full fine-tuning baseline on all six languages across all three seeds. The cross-language paired t-test rejects the null hypothesis ($t(5) = 2.948$, $p = 0.016$, $d = 1.203$), with a mean F1 gain of +1.275 points (95% CI: [0.163, 2.387]). Joint training reduces seed-to-seed variability substantially relative to single-language training, with Yoruba's standard deviation falling from 1.256 to 0.039 and Igbo's from 1.111 to 0.379. Training completes in 5.22 GPU-hours on a single NVIDIA T4 at 12.0 GB peak VRAM. These results establish that parameter-efficient fine-tuning with LoRA achieves statistically reliable improvements over full fine-tuning baselines for African NER across typologically diverse languages, that joint multilingual training with balanced sampling substantially stabilizes per-language performance across random seeds. The framework provides a scalable and computationally efficient blueprint for democratizing African NLP systems.

KEYWORDS: Parameter-efficient fine-tuning, Low-Rank Adaptation (LoRA), African language NLP, Named Entity Recognition, Cross-language equity, Multilingual transfer learning, Balanced sampling, Computational accessibility

I. INTRODUCTION

Named Entity Recognition refers to a task of recognizing and classifying entities in unstructured text. Some applications that require NER include information extraction, question answering, and knowledge graphs construction among others (Li et al., 2022). However, there have been many breakthroughs for high-resourced languages but there have been fewer developments regarding African languages, which consist of about 1.3 billion speakers belonging to over 2,000 different languages. Two interconnected challenges perpetuate this gap and form the central problem addressed in this study. First, the computational requirements of state-of-the-art multilingual NER are prohibitive for the researchers and institutions best

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

positioned to develop African language technologies. Conventional full fine-tuning of models such as XLM-RoBERTa-large demands 24–32 GB of GPU memory and 15–20 hours of training time per language (Adelani et al., 2022). Birhane et al. (2022) describes this as a form of computational exclusion that concentrates NLP research capacity in already well-resourced settings. The second problem is distributional, even where computational infrastructure is in place, it is observed that multilingual NER models exhibit systematic cross-language performance disparities. Within the MasakhaNER 2.0 benchmark, F1 scores vary by up to 14% across languages trained with the same model (Adelani et al., 2022). Alhanai et al. (2025) show that African-origin person names are recognised 15-20% less often than European names in otherwise identical contexts, pointing to biases in both training data composition and capacity allocation.

Parameter Efficient Fine-tuning (PEFT) methods, particularly LoRA (Hu et al., 2022) has cut the computational cost of adapting large language models by 90-99% on high-resource tasks, but its application to African language NER has not been studied. Comprehensive surveys in PEFT papers (Wang et al., 2025; Lialin et al., 2024) note this gap. Balanced multilingual sampling has been evaluated for European languages (Chung et al., 2023), but not in combination with parameter-efficient methods, and not on the more extreme resource disparities found in African language datasets. The study addresses both gaps by combining LoRA-based fine-tuning with balanced multilingual sampling and evaluating results across three metrics simultaneously: computational efficiency, task performance, and cross-language equity. Three objectives guided the research: (1) to establish whether LoRA achieves competitive NER performance for African languages relative to full fine-tuning baselines; (2) to quantify whether LoRA combined with balanced sampling reduces cross-language performance disparities; and (3) to validate training feasibility on consumer-grade hardware accessible globally at modest cost. The central thesis is that computational efficiency, task performance, and cross-language equity are not in fundamental tension but can be complementary objectives when parameter-efficient transfer learning is integrated with equity-aware training.

Against this background, the study makes four contributions that are, to the best of our knowledge, novel in the African NER literature. First, it provides a systematic application of LoRA to African-language NER, extending parameter-efficient fine-tuning into a typological and data-resource regime that prior PEFT literature has not evaluated (Wang et al., 2025; Lialin et al., 2024). The study used joint multi-seed evaluation and formal significance testing as methods to verify the results of each task, a methodology missing from previous studies of fine tuning NER for African languages. Secondly, we integrate balanced multilingual sampling with parameter-efficient fine-tuning; prior balanced-sampling work has been evaluated only alongside full fine-tuning or full-scale pretraining (Chung et al., 2023; Ogueji et al., 2021). Thirdly, the study demonstrated that the use of joint multilingual training can serve as both a performance enhancement and a reproducibility tool. A controlled comparison was performed comparing joint language multi-seed training versus single language multi-seed training, showing that joint training reduces seed-to-seed variance substantially. Fourth, we evaluate cross-language equity, measured by coefficient of variation, as a primary outcome alongside task performance and computational efficiency rather than as an incidental observation.

II. RELATED WORKS

A. Parameter-Efficient Fine-Tuning

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

The dominant paradigm for adapting pre-trained language models involves fine-tuning all model parameters, an approach that becomes prohibitively expensive as model sizes grow into the hundreds of billions. The landscape of parameter-efficient fine-tuning methods evolved in direct response to the growing computational cost of adapting large pre-trained models to downstream tasks. Houlsby et al. (2019) proposed inserting small bottleneck adapter modules between transformer blocks as an early PEFT alternative, though these carry inference latency costs. Pfeiffer et al. (2020) extended this to multilingual settings through AdapterHub, showing that adapter modules trained on individual languages could be composed for cross-lingual transfer without interference between language-specific representations. Prefix Tuning (Li & Liang, 2021) and Prompt Tuning (Lester et al., 2021) took a completely new direction from these methods and added learnable continuous tokens before the input to the model, while freezing the rest of the model, which reduced trainable parameters further but proved sensitive to initialisation and less stable on classification tasks than adapter-based methods.

Hu et al. (2022) reframed the adaptation problem mathematically, introducing Low-Rank Adaptation (LoRA), which approximates weight updates as low-rank matrix products $\Delta W = BA$, where $B \in R^{d\times r}$ and $A \in R^{r\times d}$ with rank $r \ll \min(m,n)$, LoRA injects trainable low-rank corrections in parallel with frozen pre-trained weights, adding no inference latency. The scaling factor used in this project ($\alpha = 2r$) follows the recommendations of Hu et al. (2021) that keeping α proportional to r produces rank-invariant update magnitudes across different experiments. Subsequent PEFT work has refined these foundations: Zhang et al. (2023) proposed AdaLoRA, which allocates rank across layers based on task importance, Paischer et al. (2025) optimised PEFT initialisation through variance analysis of pre-trained weights (EVA), obtaining faster convergence. Ding et al. (2023) confirm PEFT methods consistently approach full fine-tuning performance at substantially reduced parameter counts. Mangrulkar et al. (2022) demonstrate that retaining a fully fine-tuned classification head outside the LoRA parameterisation is critical for sequence labelling tasks. A shared limitation of this literature is that evaluations concentrate on high-resource languages and standard benchmarks, leaving open the question of whether the efficiency-performance trade-off holds under the typological diversity and data scarcity characteristic of African NLP.

B. African Language NLP and Named Entity Recognition

Research on Natural Language Processing (NLP) using African languages has progressed rapidly from 2020; however, there remain two related issues for the field to address: data scarcity and model accessibility. Joshi et al. (2020) mapped the distribution of NLP resources across the world's languages which places the majority of Africa's languages in the two most under-resourced language groups, Class 0 (No digitalized data available) and Class 1 (Only Wikipedia scale texts) and they also documented that over ninety percent of all languages worldwide (7000+) have received virtually zero attention from NLP researchers. This directly impacts model development: The XLM-RoBERTa (Conneau et al., 2020) model pre-trained on 2.5 Terabytes of filtered Common Crawl data for 100 different languages, devotes only 0.2% of its training corpus to African languages despite covering eight of them, meaning the pre-training representations for languages such as Igbo and Kinyarwanda rely on relatively thinner empirical bases compared to European languages.

Ogueji et al. (2021) questioned the idea that high resource language transfer is required in order to obtain good quality multilingual models. They trained a BERT model named AfriBERTA using only 11 low resource African languages based on less than 1 GB of text. Their approach uses batch-level language alternation and a mildly upsampled SentencePiece vocabulary distribution to ensure equitable sub-word coverage. The MasakhaNER benchmarks (Adelani et al., 2021; Adelani et al., 2022) provide the annotated data infrastructure on which both AfriBERTa and the present study rely. Adelani et al. (2021) produced the first large-scale

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

African NER benchmark (MasakhaNER) covering 10 languages. MasakhaNER 2.0 (Adelani et al., 2022) expanded coverage to 21 languages with refined annotation guidelines and native-speaker annotators. Full fine-tuning of XLM-RoBERTa-large on MasakhaNER 2.0 yielded F1 scores from 78% (Bambara) to 92% (Swahili), a 14-point spread, and all baselines used full fine-tuning without testing parameter-efficient alternatives. Alabi et al. (2025) carried out a survey synthesizing African NLP from 2020 to 2024, report that absolute performance has risen but cross-language gaps remain largely unchanged.

While the literature provides a good understanding of how the AfriBERTa model can be used with less training data than was originally available, both Ogueji et al. (2021) and Adelani et al. (2022) did not evaluate the potential use of parameter efficient fine-tuning as an adaptation technique. This study aims to address if LoRA will provide similar levels of efficiency and improvement in NER accuracy when applied to low resource languages and tasks such as those studied by the authors.

C. Multilingual Sampling Strategies and Cross-Language Equity

A persistent challenge in multilingual natural language processing is the extreme imbalance of available data in each language. In joint multilingual training, this skewed data distribution inherently biases the model's loss function toward high-resource languages, a phenomenon frequently cited as a primary catalyst for the "curse of multilinguality" (Conneau et al., 2020). Imbalance in corpus sizes among languages within the scope of multilingual NLP is a systematic bias toward the performance of machine learning models because naive random sampling leads to an unnatural distribution of samples favoring higher resource languages at the expense of accuracy of underrepresented classes (Kang and Oh, 2020). The methodological approaches used in multilingual NLP models have addressed cross-language imbalances through the use of exponential smoothing sampling (Conneau et al., 2020) to increase sample rate of lower resource languages. However, excessive repetition of lower resource corpora can lead to detrimental over fitting (Chung et al., 2023). Following from the work of Chung et al. (2023), subsequent methodologies implemented strategies that prevented extreme repetition risk (Ogueji et al., 2021), such as implementing batch level language switching (Ogueji et al., 2021) or directly limiting number of times each language may be repeated (Chung et al., 2023). Continuing with the methodology established by Ogueji et al. (2021) to apply batch-level sampling and multi-seed evaluation to low-resource African languages, the study uses a complete equalization sampling methodology. This sampling strategy raises all corpora for each target language to the same size as the largest language dataset, thus preventing the possibility of extreme repetition that Chung et al. (2023) identified as harmful.

A gap that none of these works addresses is the interaction between sampling strategy and seed variance in fine-tuning. The sampling literature focuses on pre-training or batch construction, while the reproducibility literature treats seed variability as something to be averaged-over as opposed to a variable that needs to be explained. This study's-controlled assessment of multi-seed results using both single language and joint configurations show that the use of joint training combined with balanced sampling decreases seed variance an interaction between training configuration and reproducibility previously unreported in prior studies of African NLP.

D. Research Gaps

Three critical gaps motivate the present study. First, there is no systematic study on the effectiveness of LoRA for African language NER despite extensive studies carried out on high-resource benchmarks. Second, Balanced sampling was investigated as an equity fine tuning

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

intervention in this study, however its impact on the dispersion in Cross-Language Performance has not been determined within a Controlled Ablation. Third, the idea that LoRA's parameter constraints might limit overfitting to high-resource languages, and thus reduce cross-language gaps, has been raised but not tested empirically. This study addresses all three gaps in a single experimental framework, contributing both performance evidence and methodological tools that can be applied to future African NLP work.

III. METHODOLOGY

A. Research Design

The study uses a computational experimental design. A LoRA-adapted XLM-RoBERTa-large model trained jointly on six African languages is evaluated against per-language full-fine-tuning baselines from Adelani et al. (2022). The primary measure is seqeval span-level F1. Secondary measures include precision, recall, per-entity-type F1, and the coefficient of variation across per-language F1 scores as the equity criterion. Statistical reliability is assessed via three-seed multi-seed experiments with paired bootstrap significance testing; sampling effects are isolated in a balanced vs. unbalanced ablation.

B. Dataset Preprocessing and Language Selection

The MasakhaNER 2.0 benchmark (Adelani et al., 2022) provided the evaluation data: loaded from CoNLL-format files in the masakhane-io/masakhane-ner GitHub repository; annotated for four entity types (Person, Location, Organisation, Date) in IOB2 format. Six languages were selected to maximise typological diversity within the languages available in MasakhaNER 2.0, rather than on the basis of dataset size alone:

Table 1: Dataset Sizes by Language and Split

Language	Code	Family	Train	Validation	Test
Hausa	hau	Afro-Asiatic (Chadic)	5,716	816	1,633
Yoruba	yor	Niger-Congo (Volta-Niger)	6,876	983	1,964
Swahili	swa	Niger-Congo (Bantu)	6,593	942	1,883
Igbo	ibo	Niger-Congo (Volta-Niger)	7,634	1,090	2,181
Kinyarwanda	kin	Niger-Congo (Bantu)	7,825	1,118	2,235
Nigerian Pidgin	pcm	English-lexified Creole	5,646	806	1,613
Total (raw)	—	—	40,290	5,755	11,509

These 6 languages were chosen based upon maximal typological diversity as opposed to the number of languages available with regard to their representation in the MasakhaNER 2.0. Hausa is the sole Afro-Asiatic (Chadic) language. Yoruba and Igbo are both Niger-Congo/Volta-Niger languages of West Africa but differ in morphological structure: Yoruba is tonal with limited inflectional morphology; Igbo employs a more extensive derivational system. Swahili and Kinyarwanda are Bantu languages of East and Central Africa respectively, sharing noun-class agreement but differing in geographic spread and digital text availability. Nigerian Pidgin is an English-lexified creole, introducing a structurally distinct variety with code-switching characteristics. Together the six languages represent three subfamilies and a creole, cover West, East, and Central Africa, and span a corpus-size range of 5,646 to 7,825 training sentences.

There are six different language datasets; ranging from approximately 5,646 sentences for Nigerian Pidgin to about 7,825 for Kinyarwanda. If we simply "stack" all the language data together as one large dataset for joint training it will result in languages with a greater amount of training split contributing additional gradient updates per epoch than those with a lesser number of training split. This is an example of what Kang and Oh (2020) called representation imbalance and they demonstrate how this results in biased performance assessments even

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

within a single language classification setting. Similarly, Chung et al. (2023) demonstrate how if left unchecked, this can also cause systemic disadvantages for low resource languages during multilingual pre-training.

In order to balance out any disparities in dataset sizes and to create equality among the different languages used, a balancing method was adopted: all training datasets were upsampled to match the largest language (Kinyarwanda 7,825 sentences), resulting in a total of 46,950 examples in the training sets where every language contributes an equal number of gradient updates per epoch. This approach follows Ogueji et al. (2021), who apply the same batch-level language equalisation strategy when pre-training AfriBERTa on eleven African languages and find that equal per-language representation produces more consistent cross-language NER results than proportional sampling in low-resource settings. Validation sets are balanced by the same procedure to maintain a consistent evaluation distribution; test sets remain at their original sizes to preserve comparability with the published baselines of Adelani et al. (2022).

The baseline F1 scores used in this study are drawn from Adelani et al. (2022) rather than reproduced experimentally, and several aspects of the training configuration differ between the two studies. Several elements of the comparison are confirmed identical: the test splits, since both studies use the same MasakhaNER 2.0 release and unmodified test partitions; the entity schema (PER, LOC, ORG, DATE, IOB2); and the base encoder checkpoint (XLM-RoBERTa-large). These factors establish the validity of the comparison at the level of model outputs on shared test sets, which is the standard for NLP benchmark comparisons. Several training details are not identical. . Adelani et al. (2022) report a maximum sequence length of 200 tokens against 128 tokens for this paper, a fixed 50-epoch schedule against our early-stopped 15-epoch cap, a learning rate of 5×10^{-5} against our 2×10^{-4}, and training on an NVIDIA V100 against our NVIDIA T4. The effective batch size matches exactly (their 16×2 gradient-accumulation steps and our 8×4, both totaling 32). Their paper does not report optimizer choice, learning-rate schedule, or the specific seeds used across their five averaged runs. This paper treats these differences as acceptable rather than confounding for two reasons. Full fine-tuning and LoRA have different optimal hyperparameters, so a higher learning rate for LoRA reflects standard practice rather than an unfair advantage, following the convention set by Hu et al. (2022) when introducing LoRA. On sequence length, coverage at 128 subword tokens varies substantially by language: Hausa (0.02%), Swahili (0.03%), Kinyarwanda (0.13%), and Nigerian Pidgin (0.12%) are truncated in well under 1% of training sentences, while Yoruba (11.53%, maximum length 235 tokens) and Igbo (7.64%, maximum length 145 tokens) are truncated considerably more often; pooled across all six languages, 3.46% of sentences exceed 128 tokens. If truncation at 128 tokens were a primary driver of relative performance, the two most-truncated languages would be expected to show a consistent pattern of weaker gains; instead they diverge sharply, with Igbo showing the largest improvement over baseline of any language (+3.60 F1) and Yoruba showing one of the smallest (+0.56 F1). This paper takes this as evidence that sequence-length truncation is not a dominant factor in the reported results, though it cannot be ruled out that Yoruba and Igbo performance in particular could improve further under a longer maximum sequence length, and this is flagged as a limitation in Section 5.

C. Model Architecture and Transfer Learning

The model applies transfer learning using XLM-RoBERTa-large (Conneau et al., 2020) as a frozen base encoder, with 24 transformer blocks, 16 attention heads per layer, hidden dimension 1,024, and 573,015,058 parameters total. XLM RoBERTa-large was pre-trained on 2.5 TB of filtered CommonCrawl text across 100 languages, including several African varieties. Keeping the encoder frozen preserves those multilingual representations and the volume of task-specific data required for adaptation. Low-Rank Adaptation (LoRA) adapters

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

are injected into the query, key, value, and output projection matrices of the self-attention mechanism, together with both feed-forward sub-layers, in all 24 transformer blocks. Rather than updating the full weight matrix W directly, LoRA constrains the update to a low-rank decomposition. The modified forward pass is given by Equation 1:

$$W' = W + BA \quad (1)$$

where W' is the adapted weight matrix, W is the original frozen matrix, and ΔW = BA is the low-rank update. The matrices B and A are defined by their dimensions in Equation 2:

$$W \in R^{d \times d}$$
$$B \in R^{m \times r}$$
$$A \in R^{r \times n} \quad \textbf{(2)}$$

where W ∈ Rd× d is the frozen pre-trained matrix, B ∈ $R^{m \times r}$ and A ∈ $R^{r \times n}$ are the trainable low-rank matrices, the rank r = 32 is set far below the hidden dimension d = 1024, so the product BA has at most rank 32 regardless of the values learned. This constraint limits the total number of new parameters to a small fraction of the original weight matrix. The full layer output, including the scaling factor α/r, is given by Equation 3:

$$h = W_e x + \Delta W_x = W_e x + BA_x \quad \textbf{(3)}$$

Scaled output: $h = W_e x + (\propto / r).BA_x$, where α(scaling factor)=32, r(rank)=32.

With α = 2r = 64 and r = 32, the scaling ratio α/r = 2.0, meaning the low-rank update is scaled by a factor of 2 before being added to the frozen weights. Setting α = 2r amplifies the adapter contribution, which is appropriate when adapting a general multilingual encoder to a specialised token classification task in a different linguistic domain. The total trainable parameter count follows from the number of adapted layers and the rank.

A language-agnostic NER classification head (9,216 parameters) maps the final encoder states to IOB2 entity labels. The exact set of matrices adapted, and the corresponding parameter breakdown, are detailed in Figure 1 below.

LoRA adapters were applied to all four attention projection matrices, query (W_Q), key (W_K), value (W_V), and output (W_O), together with both feed-forward sub-layers (the 1024-to-4096 intermediate projection and the 4096-to-1024 output projection), within every one of the 24 transformer blocks. Rank r = 32 was used for every adapted matrix. The query and value projections jointly determine what the model attends to and what information is propagated forward once attention is computed; the key and output projections shape the matching geometry and the recombination of attended values respectively. Including the feed-forward sub-layers extends adaptation to the component that carries the largest share of a transformer block's parameters and is most closely associated with the storage of lexical and factual associations.

Table 2: Trainable and Frozen Components of the Model

Component	Status	Params / layer	Total (×24)
Query projection (W_Q)	**Trainable (LoRA)**	65,536	1,572,864
Key projection (W_K)	**Trainable (LoRA)**	65,536	1,572,864
Value projection (W_V)	**Trainable (LoRA)**	65,536	1,572,864
Output projection (W_O)	**Trainable (LoRA)**	65,536	1,572,864
FFN intermediate (1024→4096)	**Trainable (LoRA)**	163,840	3,932,160
FFN output (4096→1024)	**Trainable (LoRA)**	163,840	3,932,160
Layer normalisation weights	Frozen	—	—

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

Token embeddings	Frozen	—	—
Position embeddings	Frozen	—	—
NER classification head	**Trainable**	—	9,225

Summing across the six adapted matrices in each of the 24 blocks gives 589,824 × 24 = 14,155,776 LoRA parameters, plus 9,225 parameters for the classification head, for a total of 14,165,001 trainable parameters out of 573,015,058, or 2.47%. Although the feed-forward sub-layers were included in this configuration, the rank-32 bottleneck still constrains each adapted matrix to a low-rank update: the LoRA adapters at the FFN layers can shift at most 32 independent directions in a 4096-dimensional space, which is a small fraction of the 4,194,304-parameter capacity of the full FFN matrix. Adaptation is therefore still tightly bounded relative to full fine-tuning, even though it spans more components than the attention-only configuration sometimes used in PEFT literature.

Only the LoRA adapters and the classification head receive gradient updates; all embeddings and layer normalisation weights remain entirely frozen throughout training.

This configuration has a more modest bearing on the equity mechanism discussed in Section 4.5 than an attention-only configuration would. Because the feed-forward layers were trainable, the model retained some capacity to adjust language-specific lexical associations. The equity improvement observed in this study is therefore better attributed to the combination of the rank-32 bottleneck, which limits how much any single matrix can shift regardless of which matrices are adapted, and the balanced sampling strategy, which equalises training exposure, rather than to an architectural exclusion of the feed-forward layers from training. Figure 1 illustrates the full framework architecture, distinguishing frozen components from trainable LoRA adapters and the fully fine-tuned classification head.

D. Training Configuration

Training used AdamW with learning rate 2×10^{-4}, and weight decay 0.01, cosine decay schedule with 10% linear warm-up. The effective batch size of 32 was obtained using 8 samples per device and 4 gradient accumulation steps. The training was allowed to run up to 15 epochs with early stopping on macro F1 score with patience of 3 epochs based on the validation dataset. The random seed was set to 42. All experiments were performed using Google Colab Pro, which provided a GPU with 16GB memory (NVIDIA T4). This setup was selected intentionally to show that the system can be accessible even to single individuals.

E. Model Training Workflow

Input sentences are tokenised using the XLM-RoBERTa SentencePiece tokeniser (Kudo & Richardson, 2018). Due to the fact that the labels used in MasakhaNER 2.0 are based on individual words rather than subwords, there must be an alignment process to map each word's BIO Label to its first subword token. All additional subword tokens and special tokens are assigned a bio label value of – 100, so that when the cross-entropy loss function is applied during either training or evaluation it will ignore them. In doing so we follow the example of Devlin et al. (2019) and avoid propagating the "B" label onto continuation tokens. Propagating

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

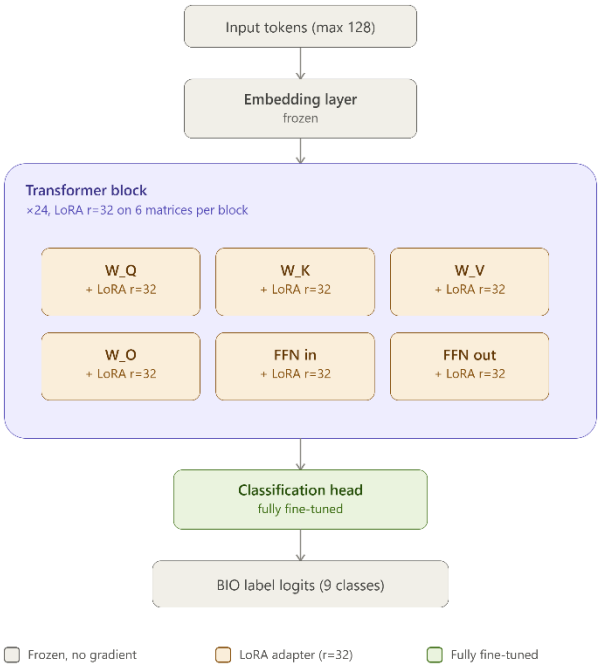

Figure 1: LoRA-based NER architecture. Frozen XLM-RoBERTa-large with LoRA adapters (r=32) on the Q, K, V and O attention projections and both feed-forward sub-layers, across all 24 transformer blocks; the classification head is fully fine-tuned.

the B- label onto continuation tokens will skew the B/I boundary distribution in morphologically complex languages like Kinyarwanda and Igbo. The sequences can then be either padded or cut off to a maximum of 128 subword tokens.

During the forward pass, the frozen XLM-RoBERTa encoder produces 1,024-dimensional contextualised representations for each token position across 24 transformer layers. At each layer, LoRA adapters modify the effective output of the W_Q, W_K, W_V, and W_O projections by adding a low-rank correction $\Delta W = BA$ ($B \in \mathbb{R}^{mx32}$, $A \in \mathbb{R}^{32xn}$) to the frozen projection output, scaled by $\alpha/r = 2$ following Hu et al. (2021). The final-layer token representations pass to the linear classification head, the only component initialised from random rather than pre-trained weights, which projects each 1,024-dimensional vector to logits over the nine BIO label classes.

The AdamW optimiser (Loshchilov & Hutter, 2019) updates only the 14.2 million parameters belonging to the LoRA matrices and the classification head; weight decay of 0.01 is applied to all trainable parameters except bias terms. The learning rate follows a cosine annealing schedule with 10% linear warm-up, peaking at 2×10^{-4} before decaying smoothly to zero. Mixed-precision training (fp16) is enabled throughout. Validation F1 is computed at the end of each epoch using the seqeval span-level protocol; training terminates if validation F1 fails to improve by at least 0.001 over three consecutive epochs, up to a maximum of 15 epochs, with the best-checkpoint weights restored for final evaluation.

Figure 2 summarises the complete training pipeline from corpus loading through per-language evaluation.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

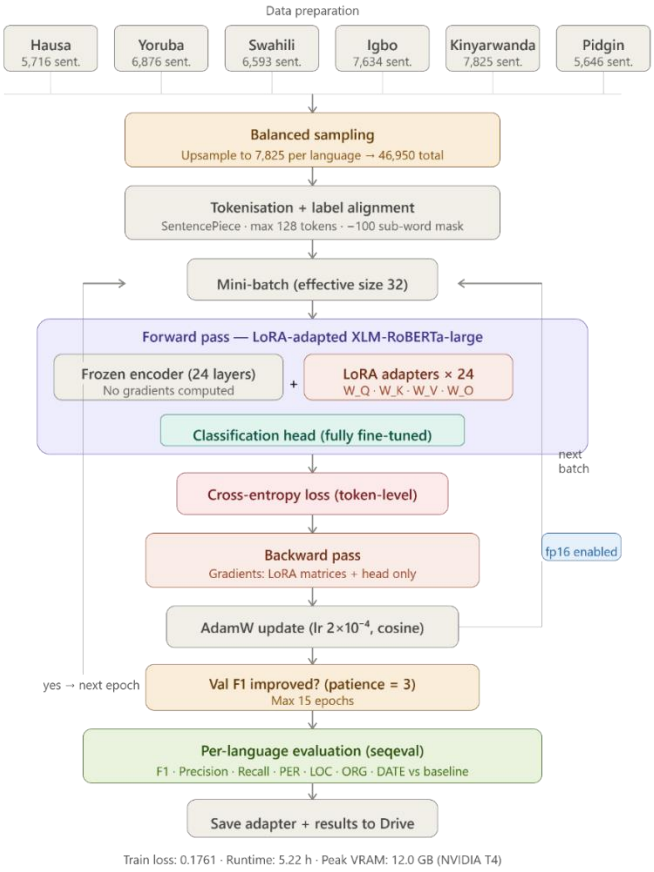

Figure 2: Training Pipeline for the LoRA-adapted XLM-RoBERTa-Large Model.

F. Evaluation Framework

Since the class imbalance in such sequential labeling datasets is quite high (with much higher counts of non-entity tokens than actual entities), raw token-level accuracy was bypassed. Rather, validation and testing steps compute metrics using the seqeval library stripping away all -100 masked placeholder values. Performance was measured using entity-level F1 scores with strict boundary and type matching following the CoNLL evaluation protocol. For each Language, Precision, Recall, and F1 are evaluated in relation to the un-modified test split; precision is a measure of the number of correctly identified spans among those spans predicted by a model, recall is a measure of the proportion of gold spans retrieved, and F1 is the harmonic mean of these two measures. To evaluate how multilingual training benefits each entity type, separate F1 scores were evaluated per entity type (PER, LOC, ORG, and DATE). The baseline evaluation uses full fine-tuned F1 scores (Adelani et al., 2022), for the same six languages on the same test splits.

The null hypothesis of the cross-language significance test is stated as H_0: $\mu\_LoRA = \mu\_baseline$, meaning that the mean F1 score for each language of the joint LoRA model will not differ from the published full fine-tuning baseline, and the one-tailed alternative is H_1: $\mu\_LoRA > \mu\_baseline$. Cross-language equity is assessed using mean F1, standard deviation, and coefficient of variation ($CV = \sigma/\mu \times 100\%$) across the six per-language F1 values, with lower CV indicating more equitable performance distribution. Statistical reliability will be evaluated by doing a joint multi-seed assessment using multiple seeds (42, 123, 2024) and reporting mean F1 and standard deviation per language alongside one-sample paired bootstrap significance tests (10,000 resamples) against each baseline. Effect size is reported as Cohen's d, and the cross-language paired t-test ($df = 5$) is used to assess whether the mean per-language

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

improvement reaches significance at $\alpha = 0.05$. To determine how much of the variation among the seeds can be attributed to jointly training the LoRA models, an additional experiment was done with the same strategy but only trained a separate LoRA model per language per seed; this allows us to compare the amount of seed-to-seed variability when we train our LoRA models separately from each other. The final experiment included evaluating the difference between using balanced sampling versus simply concatenating all of the data into a large dataset for both of our experiments (a second version of the first experiment); we used Levene's test and the F-test for variance ratio assess the statistical significance for the reduced variability in F1 cross-language scores when balanced sampling is used.

The paired t-test design is chosen over an independent-samples alternative because the six per-language baseline values vary substantially across languages; reflecting language-level difficulty rather than experimental noise and the paired structure controls for this source of variance by treating each language as its own matched unit. The one-tailed test follows directly from the directional hypothesis stated above, the study asks whether parameter-efficient fine-tuning can match or exceed full fine-tuning rather than whether it differs. Normality of the per-language mean deltas cannot be verified formally with the small sample size (six paired language-level differences); accordingly, a non-parametric Wilcoxon signed-rank test is conducted alongside the parametric paired t-test as a robustness check.

Bootstrap significance tests are conducted per language using the one-sample resampling procedure of Efron and Tibshirani (1993): each observed sample of three seed F1 values is recentred under H_0 by adding the difference between the baseline and the observed mean, 10,000 replicates of size $n = 3$ are drawn with replacement from the recentred sample. The p-value is then the proportion of replicate means at least as extreme as the observed mean in the direction of H_1. With n=3 seeds, the resampling space is limited to $3^3 = 27$ distinct outcomes, so this result should be read as corroborating rather than independently more powerful than the parametric and non-parametric tests above. The three seeds noted above were used rather than a larger number because of the compute cost of training six-language joint models within Colab Pro's session limits. Three seeds per language is the minimum feasible count given a per-run cost of 5.22 GPU-hours; the post-hoc power for the smallest observed effect ($d = 6.256$, $df = 2$, $\alpha = 0.05$, one-tailed) exceeds 0.99, though this estimate is acknowledged as circular and a minimum of five seeds is recommended for future work in this setting.

IV. RESULTS AND DISCUSSION

A. Task Performance

The entity F1 scores on a language basis using LoRA and baselines are shown in Table 3. LoRA fine-tuning showed macro-averaged F1 score improvement of 1.42% compared to full fine-tuning baseline of 87.70% while training 2.47% of the model parameters with an F1 score of 89.12%. Figure 3 visualises these results as grouped bar charts with per-language delta annotations.

Table 3: Per-language NER Performance LoRA versus Full Fine-Tuning Baseline

Language	Baseline F1	Precision	Recall	LoRA F1	Δ F1
Hausa	86.30	88.19	88.17	88.18	+1.88
Yoruba	86.60	86.34	87.99	87.16	+0.56
Swahili	92.60	91.63	94.25	92.92	+0.32
Igbo	87.20	90.70	90.90	90.80	+3.60
Kinyarwanda	84.30	84.92	86.98	85.94	+1.64
Nigerian Pidgin	89.20	91.09	88.41	89.73	+0.53
Macro average	87.70	88.81	89.45	89.12	+1.42

Performance gains in Table 3 vary by language but show no simple relationship to training set size after balanced sampling. Igbo (+3.60) and Hausa (+1.88) gained the most despite being at

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

different ends of the original size range; Swahili (+0.32) and Nigerian Pidgin (+0.53) gained least, both already near or above 89% before training. Kinyarwanda, the largest original dataset and the sampling target, gained 1.64 points, which is larger than the gains for Yoruba and Swahili despite receiving no upsampling. This suggests that the performance gain is not driven solely by the degree of upsampling, and that the pre-trained encoder's prior representations for each language play a role that the balanced sampling intervention does not fully override.

The precision-recall pattern in Table 2 is also informative. Swahili shows the largest precision-recall gap (91.63 vs 94.25), meaning the model finds most Swahili entities but produces some false positives. Kinyarwanda shows the opposite pattern (84.92 precision vs 86.98 recall), meaning the model misses some entities more often than it over-predicts them. Nigerian Pidgin shows above-average precision (91.09) with recall below the macro mean (88.41), reflecting the structural irregularity of the creole that makes boundary prediction harder than entity detection.

B. Entity Type Performance

Table 4 shows the per-entity type F1 scores across the four entity types and six languages. Two patterns stand out. The first is the ordering across entity types. PER achieves the highest F1 in five of the six languages, ranging from 91.17 (Kinyarwanda) to 98.06 (Swahili). DATE produces the lowest F1 in four of the six languages, with Igbo showing the lowest DATE score in the dataset at 72.12. The exception is Nigerian Pidgin, where DATE (90.50) is the second-highest entity type, reflecting the relatively transparent temporal expressions in that creole variety. This ranking is consistent with Adelani et al. (2022) and reflects surface-form regularity: person names appear in predictable syntactic positions and carry consistent capitalisation cues, while date expressions take diverse forms (numeric, relative, and named-period) and the DATE class typically has the fewest training tokens.

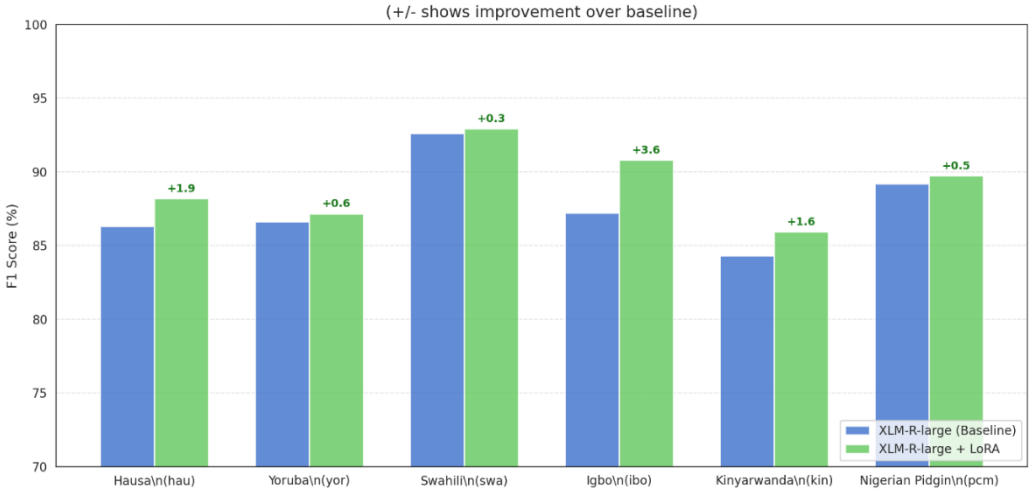

Figure 3. Per-Language F1 Scores Comparison

Table 4: Per-entity-type F1 Scores (%) by language

Language	Overall F1	PER	LOC	ORG	DATE
Hausa	88.18	96.72	88.40	85.05	79.11

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

Yoruba	87.16	91.41	88.74	81.18	82.03
Swahili	92.92	98.06	96.23	87.21	79.40
Igbo	90.80	96.84	93.79	82.45	72.12
Kinyarwanda	85.94	91.17	90.28	81.21	76.88
Nigerian Pidgin	89.73	96.39	86.53	85.78	90.50
Mean	89.12	95.10	90.66	83.81	79.99

The second pattern is within-entity variation across languages. ORG shows the narrowest per-language range, from 81.18 (Yoruba) to 87.21 (Swahili), a spread of 6.03 points. DATE shows the widest range, from 72.12 (Igbo) to 90.50 (Nigerian Pidgin), an 18.38-point spread. The wide DATE range is notable: it reflects not a uniform weakness in temporal expression recognition but structural variation in how dates are expressed across languages and text domains in MasakhaNER 2.0. Targeted data augmentation for DATE entities, or domain-specific continued pre-training with richer temporal expression coverage, would be the most direct intervention for the three languages (Hausa, Swahili, Igbo) where DATE F1 falls below 80%. Figure 4 presents these values as a heatmap; darker cells indicate higher F1 on a continuous scale from 60% to 100%.

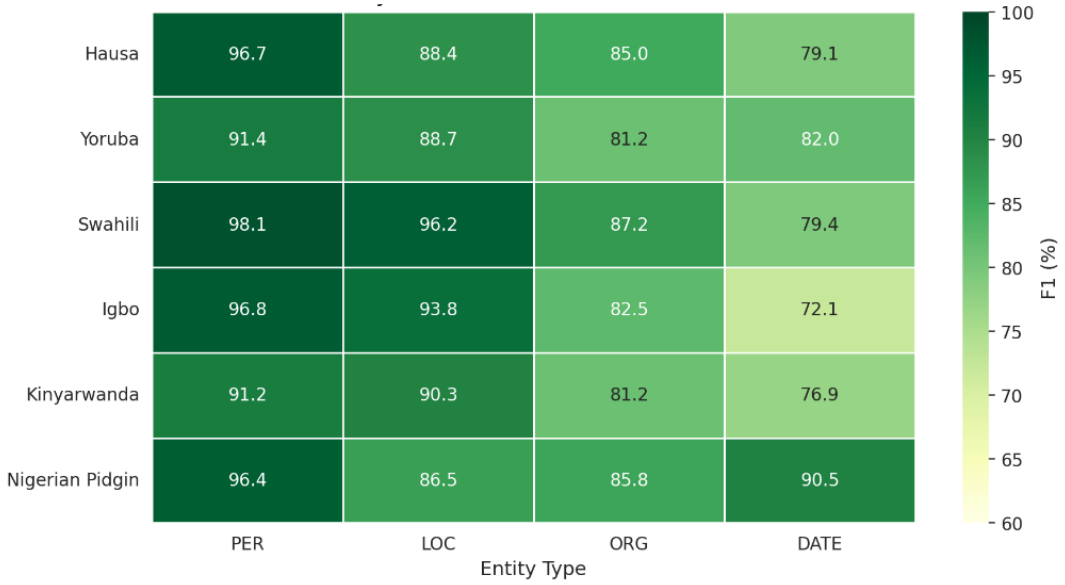

Figure 4. Per-Entity Level Heatmap

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

C. Cross-Language Equity and Sampling Ablation

Table 5 isolates the contribution of the sampling strategy to cross-language equity by placing the baseline and primary LoRA model alongside two otherwise identical joint models that differ only in whether training data is upsampled.

Table 5: Cross-Language Equity Metrics Across Conditions

Metric	Baseline (full FT)	LoRA (seed 42)	Balanced (ablation)	Unbalanced (ablation)
Mean F1 (%)	87.70	89.12	88.70	88.42
Standard deviation	2.62	2.33	2.62	2.68
Coefficient of variation (%)	2.99	2.61	2.96	3.03
F1 range (max − min)	8.30	6.98	7.57	7.55

The primary LoRA model reduces the coefficient of variation from 2.99% to 2.61% and the F1 range from 8.30 to 6.98 percentage points relative to the baseline. In the controlled sampling ablation; where the only difference is in how training data was either balanced or a concatenated dataset with no up-sampling, using a balanced distribution results in a mean F1 score of 88.70%, versus an unbalanced distribution with a mean F1 score of 88.42% (+0.28pp), and a CV reduction from 3.03% to 2.96% (a relative CV reduction of 2.32%). While there appears to be a consistent direction of the increase in equity for the single run comparison and the ablation study suggesting a true sampling effect; the existing evidence based on the number of languages tested under each condition cannot reject the null hypothesis that there will be an equivalent amount of variability in the performance of models trained on samples that are balanced versus those that are unbalanced.

D. Joint Multi-Seed Statistical Analysis

Table 6 reports mean F1 and standard deviation across three random seeds (42, 123, 2024) for the joint multilingual model evaluated per language on the MasakhaNER 2.0 test sets. Seed 42 corresponds to the primary training run reported in Section 4.1; seeds 123 and 2024 are independent replications under identical conditions. The critical one-tailed t-value at $\alpha = 0.05$ with $df = 2$ is 2.920.

Table 6: Joint Multi-seed Evaluation

Language	Seed 42	Seed 123	Seed 2024	Mean ± Std	Baseline	Δ	95% CI	Cohen's d	p (1-tail)	Sig.
Hausa	88.18	87.76	87.72	87.89 ± 0.25	86.30	+1.59	[87.26, 88.52]	6.256	0.0042	Yes
Yoruba	87.16	87.22	87.23	87.20 ± 0.04	86.60	+0.60	[87.11, 87.30]	15.574	0.0007	Yes
Swahili	92.92	92.93	92.85	92.90 ± 0.04	92.60	+0.30	[92.79, 93.01]	6.866	0.0035	Yes
Igbo	90.80	90.11	90.18	90.37 ± 0.38	87.20	+3.17	[89.42, 91.31]	8.358	0.0024	Yes
Kinyarwanda	85.94	85.58	85.69	85.74 ± 0.19	84.30	+1.44	[85.28, 86.20]	7.742	0.0028	Yes
Nigerian Pidgin	89.73	89.85	89.70	89.76 ± 0.08	89.20	+0.56	[89.56, 89.96]	6.898	0.0035	Yes

All six languages produce mean F1 values above their respective baselines across all three seeds. The improvements are statistically significant for every language ($p < 0.005$ one-tailed),

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

and the 95% confidence intervals sit entirely above the baseline in all cases — meaning the lower bound of each interval still exceeds the baseline, confirming that no language's improvement can plausibly be attributed to a single favourable initialisation. Cohen's d values range from 6.256 to 15.574, reflecting tight within-seed clustering: the three seeds produce very similar F1 values for each language, particularly for Yoruba (std = 0.039) and Nigerian Pidgin (std = 0.081), indicating that the joint model converges reliably to similar solutions regardless of initialisation.

The cross-language paired t-test yields $t(5) = 2.948$, $p = 0.016$ (one-tailed; two-tailed $p = 0.032$), with a mean delta of +1.275 F1 points (95% CI: [0.163, 2.387]) and a paired Cohen's d of 1.203. The null hypothesis that the joint LoRA model's mean per-language F1 equals the full fine-tuning baseline is rejected at $\alpha = 0.05$. As a non-parametric robustness check on this result, given the small sample size, an exact Wilcoxon signed-rank test was computed on the same six per-language deltas used above, giving $W = 21.0$, the maximum possible value at $n = 6$, with a one-tailed $p = 0.016$ and two-tailed $p = 0.0312$. This corroborates the paired t-test result without relying on a normality assumption.

These results contrast sharply with a parallel single-language multi-seed analysis, in which each language was fine-tuned independently on its own corpus across the same three seeds. Under single-language training, Yoruba and Igbo showed statistically significant under-performance relative to the baseline (Yoruba: $\Delta = -2.17$, $p = 0.048$; Igbo: $\Delta = -2.86$, $p = 0.023$), and the cross-language paired t-test failed to reject the null hypothesis ($t(5) = -1.337$, $p = 0.119$). The divergence between the two analyses is attributable to the regularising effect of joint training: when a model learns from six languages simultaneously, the shared adapter parameters are constrained by all six language distributions, reducing sensitivity to random initialisation for any individual language.

Table 7: Variance Comparison between Joint vs. Single-Language Multi-Seed Training

Language	Joint std	Single-language std	Difference
Hausa	0.254	0.501	−0.247
Yoruba	0.039	1.256	−1.217
Swahili	0.043	0.263	−0.220
Igbo	0.379	1.111	−0.732
Kinyarwanda	0.186	0.381	−0.195
Nigerian Pidgin	0.081	0.381	−0.300

Joint std: standard deviation of per-language F1 across three seeds under joint multilingual training. Single-language std: standard deviation under independent per-language training. Negative difference indicates lower variability under joint training.

This consistent improvement in variance is seen in all six languages without any exception. The largest improvement is seen in Yoruba (−1.217) and Igbo (−0.732), being the two languages that are most sensitive to seeds when fine-tuned under single-language training. Instead of just filtering out noise, multilingual joint training helps converge stability more effectively for languages where single-language fine-tuning is inconsistent. For benchmarking purposes, the takeaway from this is simple – single-language fine-tuning shows greater seed-to-seed inconsistency across runs of this dataset size, while multilingual joint training shows a consistent reduction in that variability across all six languages tested, though confirming this pattern with a larger number of seeds remains an important direction for future work.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

Three seeds are sufficient to show a consistent direction of effect across all six languages, but they remain a small sample for characterising reproducibility with statistical rigor. The standard deviations in Table 6 and Table 7 should be read as indicative rather than precise population estimates. Wider seed sampling, ten or more seeds, would be needed to establish tighter confidence bounds on seed-to-seed variance itself. The study presents the joint-versus-single-language variance comparison as suggestive evidence of a regularisation effect rather than a fully established reproducibility guarantee, and encourage replication with larger seed counts in future work.

E. Computational Efficiency

The approach demonstrated compelling efficiency advantages across multiple dimensions. The training completed in 5.22 hours with the maximum GPU memory consumption of 12.0 GB using the NVIDIA T4. The reduction from 573 million to 14.2 million trainable parameters further enables adapter-only distribution: LoRA weights can be stored and shared independently of the frozen backbone, permitting lightweight version control and multi-language adapter swapping on a single shared base model without the storage overhead of independently fine-tuned full models.

F. Confusion Matrix Analysis

The aggregate confusion matrix in figure 4 pools predictions across all 306,115 test tokens from six languages. The O class achieves 99.1% token-level recall, confirming the model is conservative about assigning entity labels to non-entity tokens. Among entity classes, B-PER and I-PER achieve recalls of 96.6% and 98.3% respectively.

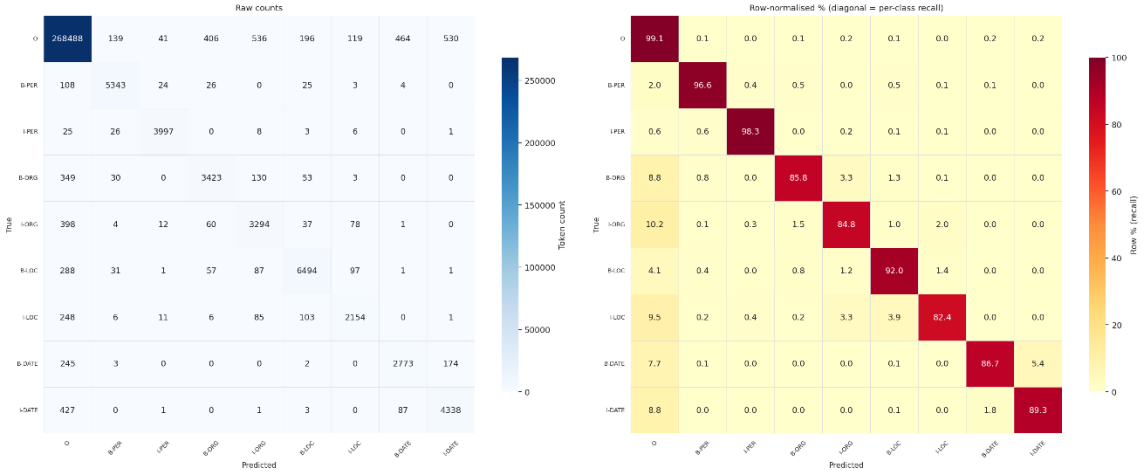

Figure 4: *Aggregate confusion matrix — all labels, all six languages combined (306,115 test tokens). Left panel: raw token counts. Right panel: row-normalised percentages (diagonal = per-class recall). The dominant off-diagonal errors are B-ORG predicted as B-LOC (1.5%) and I-LOC predicted as B-LOC (4.4%), the latter representing boundary errors at the start of multi-token location spans.*

In the entity-only aggregate matrix in figure 6 the most common cross-type confusions are B-ORG predicted as B-LOC (1.5% of true B-ORG tokens) and I-LOC predicted as B-LOC (4.4%), the latter representing boundary errors at the continuation of multi-token location spans.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

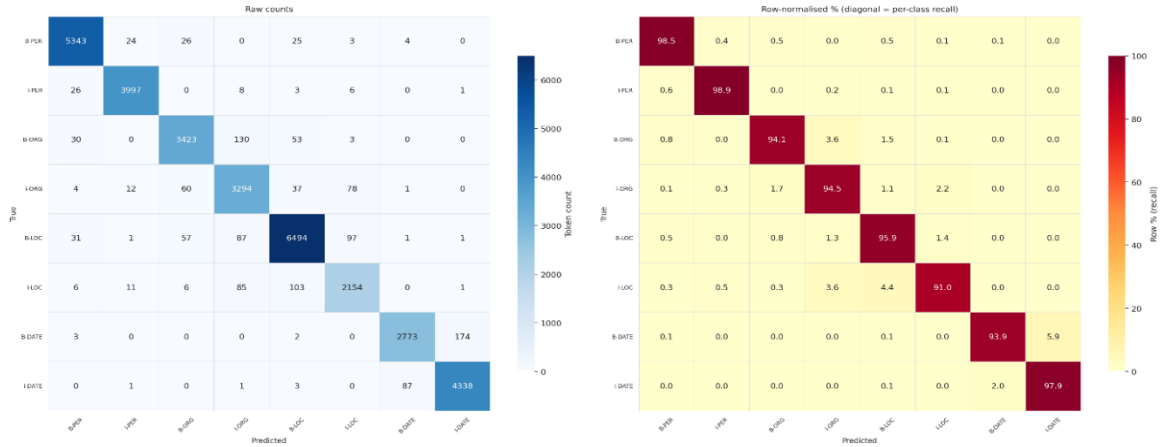

Figure 5: *Aggregate confusion matrix — entity classes only (O excluded). Row-normalised percentages reveal cross-type confusion patterns independent of the O-class dominance. B-ORG recall is 94.1% and I-ORG recall is 94.5%; the primary error for both is false negatives absorbed into the O class when the full label set is included (visible in Figure 4). I-LOC achieves 91.0% recall with 4.4% of tokens predicted as B-LOC*

G. Discussion and Limitations

All these findings together confirm the thesis that efficiency, performance, and fairness can be achieved at simultaneously through the synergy of transfer learning, which gives multilingual basics to reduce the data needed for task-specific adaptation; balanced sampling, which provides the same training exposure across each language; and LoRA parameter constraints, which prohibit memorization of upsampled samples in the model and force cross-lingual generalization. The equity results are directionally consistent with balanced sampling contributing to reduced cross-language variance (coefficient of variation: 2.99% to 2.61% for the primary comparison, 3.03% to 2.96% in the controlled ablation). However, no statistically significance using either Levene's test or the F-test for variance ratio in the six-language ablation. Hence, this is treated as suggestive rather than confirmatory evidence and do not claim that balanced sampling causally reduces cross-language disparity on the basis of this study alone.

The current design confounds three interventions: multilingual joint training, balanced sampling, and the LoRA parameter constraint. This makes it difficult to attribute the observed 89.12% F1 and equity gains (coefficient of variation: 2.99% to 2.61%) to any single component. Table 5 can be used to isolate some of the sampling effects, as it compares the results of the two different methods for training the model jointly with LoRA, in both cases the same architecture and the same joint training design were used; however, it does not isolate the LoRA constraint itself from full fine-tuning under otherwise identical joint, balanced training. A fully disentangling design would need a 2×2 factorial: full fine-tuning versus LoRA, crossed with balanced versus unbalanced sampling, both evaluated under joint multilingual training. This would allow for the main effect of parameter efficiency and the main effect of sampling strategy be estimated independently, along with their interaction. Carrying out the full fine-tuning arm of this factorial was not feasible within the compute budget available for this study, since full fine-tuning of XLM-RoBERTa-large across six languages jointly would require substantially more GPU-hours and VRAM than the T4 used here; this study flags this as the clearest direction for follow-up work.

The study is subject to several limitations. First, the full fine-tuning baseline is drawn from published results (Adelani et al., 2022) rather than reproduced independently. The test splits, entity schema, base encoder, and effective batch size are confirmed identical, but maximum sequence length, epoch budget, learning rate, and hardware differ, and their optimizer and schedule are not disclosed. Second, the multi-seed reproducibility analysis utilized 3 random

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

seeds. All 6 languages showed a consistent direction of reduced seed variance after joint training; however, three seeds cannot establish precise reproducibility bounds, and hence, a larger seed count is recommended for future work. Third, the balanced-sampling ablation shows a directionally consistent but not statistically significant reduction in cross-language variance (Levene's test and F-test both non-significant at six languages), and should be read as suggestive rather than confirmatory. Fourth, the current design confounded joint multilingual training, balanced sampling and the LoRA parameter constraint; a full factorial design, crossing full fine-tuning versus LoRA with balanced versus unbalanced sampling, would be needed to isolate their individual contributions, but was not feasible within the available compute budget.

V. CONCLUSION

This study set out to determine whether parameter-efficient fine-tuning via LoRA could match or exceed full fine-tuning baselines for named entity recognition across six typologically diverse African languages from MasakhaNER 2.0: Hausa, Yoruba, Swahili, Igbo, Kinyarwanda, and Nigerian Pidgin; while remaining within the computational constraints of consumer grade hardware and reducing cross-language performance disparity. The performance and efficiency objectives are met. The jointly-trained LoRA adapter, updating 2.47% of XLM-RoBERTa-large's parameters, surpasses the published full fine-tuning baseline on all six languages across all three evaluation seeds, with the cross-language paired t-test rejecting the null hypothesis ($t(5) = 2.948$, $p = 0.016$, $d = 1.203$) and individual improvements significant at $p < 0.005$ for every language. The complete training run completes in 5.22 GPU-hours on a single NVIDIA T4 at 12.0 GB peak VRAM; a profile compatible with standard academic cloud allocations and the adapter weights are small enough to distribute independently of the frozen backbone, lowering the barrier to deployment and model sharing in low-resource settings. The equity objective is supported directionally rather than confirmed at conventional levels of statistical significance: the coefficient of variation across languages fell from 2.99% under full fine-tuning to 2.61% under the joint LoRA model, but this reduction was not established as statistically significant in the six-language balanced-versus-unbalanced ablation reported in Section 4.3.

In addition to demonstrating strong performance results, this study also has two significant contributions in terms of methodology for the broader African NLP community. First, we see how comparing the seed-to-seed variance using single-language and joint multilingual training, shows that when we train models together in a multilingual fashion, the average variability among seeds is reduced significantly: while averages of 0.65 F1 points/seed/language were seen using single-language training, we find averages of less than 0.16 F1 points/seed/language when training using joint-multilingual fine-tuning, and this was seen in each of the six languages. As noted in Section 4.4, three seeds are sufficient to establish a consistent direction for this effect but not to establish precise reproducibility bounds, and replication with a larger number of seeds would strengthen this finding. Therefore, instead of seeing joint-training solely as a way to improve performance, we can now use it as an additional way to reduce variability among seeds (and therefore improve reproducibility), and thereby indicate that the size of the dataset used to evaluate model quality through benchmarking (approximately 5,000-8,000 sentences) will be too small to make reliable inferences regarding model quality based on single-run, single-language evaluations. Second, although the ablation test examining whether or not having a balanced representation of the population (as defined by demographic factors) improved equity in NER evaluation did not produce a statistically-significant result at six conditions/languages, the reduction in coefficients of variation from 2.99% to 2.61% is directionally consistent with equity improvement; further analysis and experimentation using larger numbers of languages would provide stronger evidence. Pragmatically speaking, the empirical studies support recommendations for African NLP researchers who have limited

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21540914

resources (i.e., limited computing capacity) to prioritize training their NER models jointly across multiple languages rather than doing separate fine-tunings for each individual language; and to report results for all seeds run for each training condition rather than just running a single instance of each condition. Finally, these results support treating the training configurations themselves as variables that have empirically demonstrable effects on both performance and reproducibility.

REFERENCES

Adelani, D.I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I.A., Muhammad, S.H., Emezue, C.C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S.M., Gwadabe, T.R., Ezeani, I., Niyongabo, R.A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E.P., Oyerinde, S., Siro, C., Bateesa, T.S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B.F.P., Ogueji, K., DIOP, T.I., Diallo, A., Akinfaderin, A., Marengereke, T. & Osei, S. (2021). MasakhaNER: named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9, 1116–1131.

Adelani, D.I., Neubig, G., Ruder, S., Rijhwani, S., Beukman, M., Palen-Michel, C., Lignos, C., Alabi, J.O., Muhammad, S.H., Nabende, P., Dione, C.M.B., Bukula, A., Mabuya, R., Dossou, B.F.P., Sibanda, B., Buzaaba, H., Mukiibi, J., Kalipe, G., Mbaye, D., Taylor, A., Kabore, F., Emezue, C.C., Aremu, A., Ogayo, P., Gitau, C., Munkoh-Buabeng, E., Koagne, V.M., Tapo, A.A., Macucwa, T., Marivate, V., Mboning, E., Gwadabe, T., Adewumi, T., Ahia, O., Nakatumba-Nabende, J., Mokono, N.L., Ezeani, I., Chukwuneke, C., Adeyemi, M., Hacheme, G.Q., Abdulmumin, I., Ogundepo, O., Yousuf, O., Ngoli, T.M. & Klakow, D. (2022). MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. *Proceedings of EMNLP 2022*.

Alabi, J. O., Hedderich, M. A., Adelani, D. I., & Klakow, D. (2025). Charting the Landscape of African NLP: Mapping Progress and Shaping the Road Ahead. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 27795 27829. https://doi.org/10.18653/v1/2025.emnlp-main.1414

Alhanai, T., Kasumovic, A., Ghassemi, M. M., Zitzelberger, A., Lundin, J. M., & Chabot Couture, G. (2025). Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages with New Benchmarks, Fine-Tuning, and Cultural Adjustments. Proceedings of the AAAI Conference on Artificial Intelligence, 39(27), 27802–27812.

Birhane, A., Kasirzadeh, A., Leslie, D. & Wachter, S. (2022). Science in the age of large language models. Nature Reviews Physics, 4(12), 761-762.

Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic Inequalities in Language Technology Performance across the World's Languages. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 5486–5505. https://doi.org/10.18653/v1/2022.acl-long.376

Chung, H. W., Constant, N., Garcia, X., Roberts, A., Tay, Y., Narang, S., & Firat, O. (2023). Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*.

Conneau, A., Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H., Chen, J., Liu, Y., Tang, J., Li, J. & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 5(3), 220-235.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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)* (pp. 4171-4186).

Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability, 57. New York: Chapman & Hall.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M. & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Proceedings of the 36th international conference on machine learning (2790-2799).

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. (2022). LoRA: low-rank adaptation of large language models. In Proceedings of the international conference on learning representations (ICLR).

Kang, D., & Oh, S. (2020). Balanced training/test set sampling for proper evaluation of classification models. *Intelligent Data Analysis*, *24*(1), 5-18.

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer. *Proceedings of EMNLP 2018 System Demonstrations*, 66–71. https://aclanthology.org/D18-2012

Joshi, P., Santy, S., Budhiraja, A., Bali, K. & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293.

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *Proceedings of EMNLP 2021*, 3045–3059. https://aclanthology.org/2021.emnlp-main.243

Li, J., Sun, A., Han, J. & Li, C. (2022). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70.

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of ACL-IJCNLP 2021*, 4582–4597. https://aclanthology.org/2021.acl-long.353

Lialin, V., Deshpande, V., Yao, X. & Rumshisky, A. (2024). Scaling down to scale up: a guide to parameter-efficient fine-tuning. arXiv preprint arXiv:2303.15647.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., & Bossan, B. (2022). Peft: State-of-the-art parameter-efficient fine-tuning methods.

Ogueji, K., Zhu, Y., & Lin, J. (2021, November). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st workshop on multilingual representation learning* (pp. 116-126).

Paischer, F., Hauzenberger, L., Schmied, T., Alkin, B., Deisenroth, M.P. & Hochreiter, S. (2025). Parameter efficient fine-tuning via explained variance adaptation (EVA). In Proceedings of the 39th conference on neural information processing systems (NeurIPS 2025).

Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020, November). Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 7654-7673).

Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., Ding, N., Liu, Z. & Sun, M. (2025). Parameter-efficient fine-tuning in large language models: a survey of methodologies. Artificial Intelligence Review, 58, 227.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W. & Zhao, T. (2023). AdaLoRA: adaptive budget allocation for parameter-efficient fine-tuning. In Proceedings of the international conference on learning representations (ICLR).