Rethinking Full Finetuning from Pretraining Checkpoints in Active Learning for African Languages

Bonaventure F. P. Dossou[1,2], **Inès Arous**[3*], **Jackie Chi Kit Cheung**[1,2,4]

[1] McGill University [2] Mila Quebec AI Institute [3] York University, [4]Canada CIFAR AI Chair
{bonaventure.dossou, cheungja}@mila.quebec, inesar@york.ca

Abstract

Active learning (AL) aims to reduce annotation effort by iteratively selecting the most informative samples for labeling. The dominant strategy in AL involves fully finetuning the model on all acquired data after each round, which is computationally expensive in multilingual and low-resource settings. This paper investigates *continual finetuning* (CF), an alternative update strategy where the model is updated only on newly acquired samples at each round. We evaluate CF against full finetuning (FA) across 28 African languages using MasakhaNEWS and SIB-200. Our analysis reveals three key findings. First, CF matches or outperforms FA for languages included in the model's pretraining, achieving up to 35% reductions in GPU memory, FLOPs, and training time. Second, CF performs comparably even for languages not seen during pretraining when they are typologically similar to those that were. Third, CF's effectiveness depends critically on uncertainty-based acquisition; without it, performance deteriorates significantly. While FA remains preferable for some low-resource languages, the overall results establish CF as a robust, cost-efficient alternative for active learning in multilingual NLP. These findings motivate the development of hybrid AL strategies that adapt fine-tuning behavior based on pretraining coverage, language typology, and acquisition dynamics. Our code is available here.

1 Introduction

Building effective NLP systems for low-resource languages requires strategies to optimize the use of limited data and infrastructure. Active learning (AL) offers a compelling solution by focusing annotation efforts on the most informative samples, thereby maximizing model performance under tight resource constraints (Dossou et al., 2022). This is especially critical for African languages, where labeled corpora are expensive to collect and often unavailable. Uncertainty-based acquisition methods such as Monte Carlo Dropout (Gal and Ghahramani, 2016; Gal et al., 2017a), BALD (Gal et al., 2017b), and BatchBALD (Kirsch et al., 2019) have been shown to reduce labeling needs while maintaining accuracy. These techniques make AL particularly suited to multilingual NLP in data-scarce contexts (Settles, 2012; Lewis and Gale, 1994; Cohn et al., 1996). Yet, computational resources are also constrained in many of these same settings, making it equally important to consider the cost of model updates during training and the costs associated with annotation.

The standard practice in AL is to fully finetune from scratch or pretraining checkpoints at each acquisition round, using all accumulated labeled data. While this approach has proven effective, it becomes computationally expensive as the dataset grows, requiring more GPU memory and longer training time (Dossou et al., 2022; Gal et al., 2017b; Kirsch et al., 2019). Given the rising computational demands of large-scale models (Grattafiori et al., 2024; Chowdhery et al., 2022; Hoffmann et al., 2022; Kaplan et al., 2020; Patterson et al., 2021), we investigate the following research questions: how can both computational and annotation costs in AL frameworks be balanced without compromising effectiveness? Instead of fully finetuning on all accumulated data, could updating the model solely on newly acquired samples provide a more computationally efficient alternative? To answer this, we explore continual finetuning, where the model is incrementally updated using only newly acquired samples at each AL round.

In this paper, we conduct experiments on MasakhaNEWS (Adelani et al., 2023b) and SIB-200 (Adelani et al., 2023a), two datasets covering multiple African languages. We compare two AL finetuning strategies: (1) finetuning from pre-

*This work was done while the author was at Mila and McGill University.

training checkpoints on all acquired data and (2) continual finetuning solely on newly acquired samples. Our evaluation examines whether the latter maintains model performance while reducing computational costs. Our study aims to provide insights into the trade-off between computational and annotation costs in active learning.

Our results show that continual finetuning reduces GPU memory usage by 30–35%, FLOPs by 32–38%, and clock time by 35–40%, significantly lowering computational costs. In terms of performance, continual finetuning achieves comparable and even better performance when languages are part of the pretraining corpus. However, for underrepresented languages not part of the pretraining corpus, full finetuning helps the model integrate new information effectively and mitigates instability of downstream performance caused by distributional shifts. These findings challenge the assumption that AL must always involve full finetuning on all acquired data and highlight trade-offs between computational costs and model performance.

Our main contributions are: **(1)** we present the first comparative study of full versus continual finetuning in active learning, across 28 African Languages; **(2)** we quantify the computational saving of continual finetuning in terms of memory usage, FLOPS, and wall-clock time; **(3)** we analyze performance trends across languages seen and unseen during pretraining, revealing when continual finetuning is sufficient or insufficient; **(4)** we challenge the common assumption that full finetuning is necessary at each acquisition round in active learning, offering practical alternatives for low-resources languages.

2 Related Work

2.1 Active Learning in NLP

Active learning (AL) is widely used in NLP to reduce annotation costs by selecting the most informative samples for labeling (Settles, 2012; Lewis and Gale, 1994; Cohn et al., 1996). Most work focuses on acquisition strategies, including uncertainty-based methods like MC Dropout (Gal and Ghahramani, 2016), BALD (Houlsby et al., 2011), and CoreSet (Sener and Savarese, 2018), which have proven effective for tasks such as classification and sequence labeling (Ein-Dor et al., 2020; Maekawa et al., 2022; Schröder et al., 2022; Hübotter et al., 2024). However, this literature emphasizes annotation cost while largely overlooking the growing computational demands of retraining large models (Hoi et al., 2006; Kirsch et al., 2023; Azimi et al., 2012; Guo and Schuurmans, 2008). Many studies assume full retraining after each round (Gal et al., 2017b; Dossou et al., 2022; Kirsch et al., 2019, 2023), an approach that is impractical in low-resource settings where compute access is also constrained (Dossou et al., 2022; Dossou, 2023). Our work revisits this assumption and isolates the role of update strategies, offering a new perspective that accounts for both annotation and computational costs.

2.2 African Languages in NLP

African languages are underrepresented in NLP due to limited labeled data, low digital presence, and scarce pretraining coverage (Nekoto et al., 2020; Dossou et al., 2022). These languages belong to families such as Bantu (e.g., Zulu, Xhosa), Afro-Asiatic (e.g., Amharic, Hausa), and Niger-Congo (e.g., Yoruba, Fon), and exhibit diverse characteristics in terms of tone, morphology, and script. Some, such as Swahili and Hausa, have moderate coverage, while others remain extremely low-resource languages. Benchmarks such as MasakhaNEWS (Adelani et al., 2023b) and SIB-200 (Adelani et al., 2023a) have helped advance the field, but core ML research still rarely explores methodological choices that reflect the realities of African NLP. Our work addresses this by evaluating continual finetuning across 28 African languages, analyzing how typology, pretraining, and acquisition strategy interact in active learning.

2.3 Continual Finetuning and Links to Continual Learning

Continual finetuning (CF) updates models only on newly acquired samples, rather than all labeled data, thereby reducing memory usage, floating-point operations (FLOPs), and runtime. Though CF has been studied in multi-task and domain adaptation (Aggarwal et al., 2024; Mundt et al., 2023; Ayub and Fendley, 2022), little work has examined its role in AL, particularly for diverse or multilingual settings. Broader continual learning (CL) focuses on incremental updates and preventing forgetting across tasks (Parisi et al., 2019), often using memory or regularization techniques (Das et al., 2023). Our approach is intentionally simple: an architecture-agnostic CF strategy that avoids CL-specific modifications. We aim to assess whether this lightweight alternative can match full retrain-

ing in AL, especially in resource-constrained multilingual environments.

3 Experimental Setup

This section outlines our experimental protocol for evaluating active learning (AL) update strategies in multilingual, low-resource African natural language processing (NLP) settings. We describe the AL framework and sampling strategy, detail the datasets and models used, and explain our evaluation metrics and computational budget.

3.1 Active Learning Strategies

Our active learning (AL) setup follows a standard iterative pipeline. Given an initial labeled dataset $\mathcal{D}_{\text{train}}$ and an unlabeled pool $\mathcal{U}$, AL proceeds in rounds as follows:

1. Train the model f_θ on the current labeled dataset $\mathcal{D}_{\text{train}}$
2. Use an acquisition function to select a batch $\mathcal{Q}_{r'} \subset \mathcal{U}$ of unlabeled samples
3. Annotate $\mathcal{Q}_{r'}$ and update the labeled set: $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \mathcal{Q}_{r'}$
4. Update the model

We compare two update strategies: (1) *Finetuning All (FA)*, where the model is retrained from the original pretraining checkpoint on the full labeled dataset after each round, and (2) *Continual Finetuning (CF)*, where the model is updated only on the most recently acquired batch $\mathcal{Q}_{r'}$. This process repeats for $r = 10$ rounds or until the pool $\mathcal{U}$ is exhausted.

We use uncertainty sampling with Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) for sample acquisition. Specifically, we perform 10 stochastic forward passes with dropout enabled at inference time. We compute the average token-level entropy for each sample in $\mathcal{U}$ and select the top 100 most uncertain examples to be labeled and added to the training set. This method ensures the model prioritizes informative or ambiguous instances.

3.2 Datasets and Model

We evaluate our setup using two African NLP benchmarks: **MasakhaNEWS** (Adelani et al., 2023b) and **SIB-200** (Adelani et al., 2023a), both designed to support evaluation in multilingual, low-resource, and typologically diverse settings.

MasakhaNEWS is the largest human-annotated dataset for multilingual news classification in African languages. It spans **16 languages** from across Africa and includes **7 topic labels** (e.g., politics, health, sports). Articles were sourced from trusted outlets, such as the *BBC* and *VOA*, with document counts per language ranging from 1,000 to over 10,000. Annotation was performed in two stages by native speakers using active learning, yielding Fleiss Kappa scores ranging from 0.55 to 0.85.

SIB-200 is a sentence-level classification dataset derived from Flores-200. It includes **1,004 annotated examples across 205 languages and dialects**, covering 21 African language families such as Bantu, Afro-Asiatic, Nilotic, and Mande. The data spans seven topics, offering broad typological and domain diversity for evaluating multilingual models.

We use the official train/validation/test splits for all experiments. As our base model, we adopt **AfroXLMR-Large** (Alabi et al., 2022), a multilingual encoder-only Transformer derived from XLM-RoBERTa, finetuned on 17 African languages. AfroXLMR is favored for its open-source nature, classification compatibility, and efficiency, in contrast to decoder-only LLMs like GPT (Brown et al., 2020), Gemini (Team et al., 2023), or LLaMA (Grattafiori et al., 2024). While newer models such as Aya (Üstün et al., 2024) are emerging, AfroXLMR remains a robust and practical choice for African NLP.

All experiments are run on two NVIDIA A100 GPUs (each with 48GB VRAM and 6 CPU cores), with a maximum runtime of 10 hours. We perform 10 active learning rounds, acquiring 100 new samples per round. Full hyperparameter settings are provided in Table 4.

3.3 Evaluation Metrics

We evaluate model performance using the mean F1 score across all AL rounds, a standard metric for summarizing acquisition effectiveness (Gal et al., 2017b; Kirsch et al., 2019; Jain et al., 2023). We also compute the standard deviation of F1 scores to assess performance stability over time. Full per-round trends are visualized in Figures 2 and 3. We track GPU memory usage, floating point operations (FLOPs), and wall-clock time in hours to assess efficiency. FLOPs are computed using the `fvcore` PyTorch utility. These measurements allow us to quantify the trade-off between computational cost and predictive performance across update strategies.

4 Results and Analysis

This section presents empirical findings on the effectiveness of Continual Finetuning (CF) compared to Finetuning All (FA) across multiple African languages using active learning. Our results are organized around three key findings: (1) languages included in the pretraining corpus of the model benefit most from CF; (2) linguistic proximity to pretraining languages improves outcomes; and (3) principled sample selection strategies are critical for CF's success. We conclude each finding by discussing its implications for selecting the optimal update strategy in multilingual AL settings.

4.1 Finding 1: Languages Covered During Pretraining Benefit Most from Continual Finetuning

Languages included in the pretraining corpus of AfroXLMR consistently benefit from CF. As shown in Figure 1, CF matches or outperforms FA for languages such as Yoruba (yor), Swahili (swa), and Hausa (hau) in MasakhaNEWS, and Sesotho (sot), Afrikaans (afr), Zulu (zul), and Xhosa (xho) in SIB-200. These languages benefit from both strong initial representations and, in the case of MasakhaNEWS, relatively larger training sample sizes, which likely contribute to stable learning under CF.

CF also achieves significant resource savings: GPU memory usage, FLOPs, and training time are reduced by 33.56%, 33.78%, and 34.83%, respectively, in MasakhaNEWS, with similarly large savings in SIB-200 (Tables 1, 2). These gains are significant for multilingual active learning, where repeated model updates can be prohibitively expensive.

To assess whether the performance differences between CF and FA are statistically meaningful, we apply the Wilcoxon signed-rank test, a non-parametric method used to evaluate the significance of paired differences across rounds. Results in Table 3 confirm that CF is a competitive alternative to FA. In SIB-200, no language shows a statistically significant difference between CF and FA across active learning rounds. In MasakhaNEWS, 9 out of 14 languages show substantial differences that favor FA. However, the corresponding effect sizes are usually small or negligible, indicating limited practical relevance. These results suggest that CF offers a compelling trade-off between computational efficiency and predictive performance for languages covered during pretraining.

4.2 Finding 2: Linguistic Proximity Amplifies Continual Finetuning Success

CF also performs well for languages not explicitly included in pretraining but closely related to those that are. In both datasets, several Bantu languages such as Luganda (lug), Tswana (tsn), Tsonga (tso), and Luo (luo) benefit from CF despite not being part of AfroXLMR's pretraining. These languages belong to the Niger-Congo phylum, specifically the Bantu family, which includes pretraining languages like Zulu (zul) and Xhosa (xho).

Per-round performance curves (Figures 2 and 3) show that Bantu languages typically exhibit smoother and more stable trajectories under CF. This is likely due to shared linguistic features such as noun class systems, agglutinative morphology, and common syntactic structures. These patterns suggest that linguistic similarity allows CF to generalize effectively across typologically related languages without explicit pretraining.

In contrast, Afro-Asiatic languages such as Amharic (amh), Tigrinya (tir), and Hausa (hau) show greater volatility under both CF and FA. These languages are typologically distant from the Bantu family and possess unique orthographic and morphosyntactic characteristics. For instance, Amharic and Tigrinya use the Ge'ez script, which is not observed in any other training languages, and they are low-resource even within their own family. FA tends to perform better for these languages, particularly in later rounds, possibly because full updates allow the model to incorporate more task-specific structural information gradually.

West African Niger-Congo languages such as Yoruba (yor), Igbo (ibo), Fon (fon), and Ewe (ewe) show mixed results. While Yoruba consistently benefits from CF, others like Fon and Ewe experience erratic performance. This likely results from inconsistent lexical overlap, limited dataset quality, or insufficient pretraining exposure. This variability highlights the limitations of generalizing solely from language family and emphasizes the importance of resource quality and orthographic alignment.

These patterns align with the findings of Adelani et al. (2022), who show that genetic, syntactic, and phonological similarity among African languages correlates with transfer effectiveness in multilingual models. Based on family classification, phoneme inventory overlap, and syntactic tem-

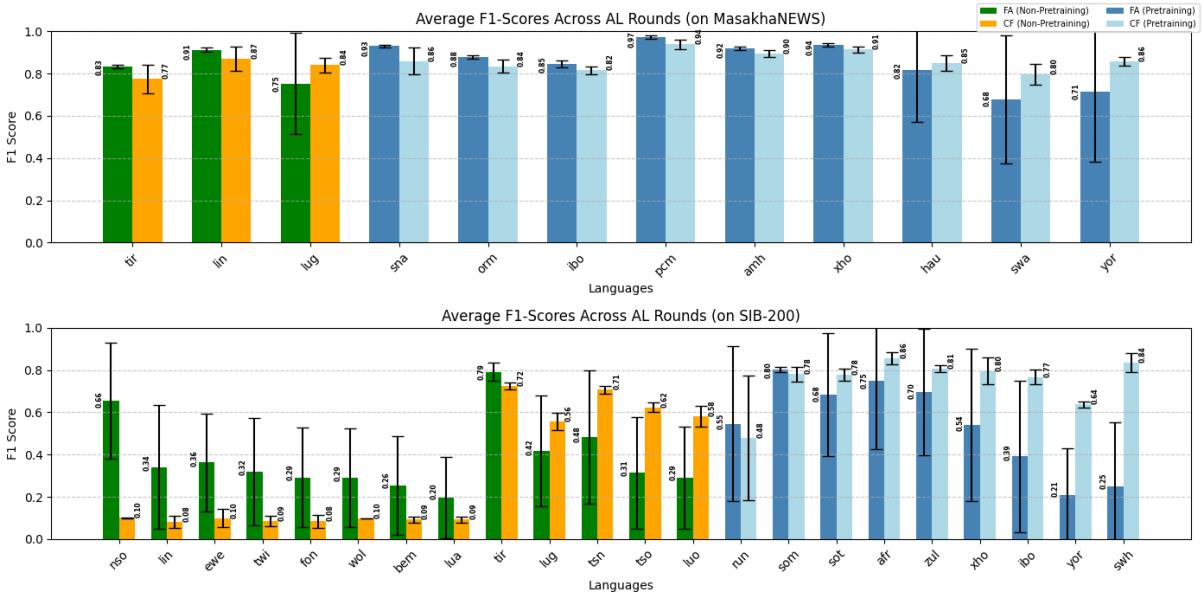

Figure 1: Average F1-Scores Across AL rounds for each language in MasakhaNEWS and SIB-200, using **FA** and **CF**. **Pretraining/Non-Pretraining** indicates whether the language was included in the pretraining set of the AfroXLMR-Large model. Within each group (Pretraining, Non-Pretraining), languages are sorted based on the percentage improvement of **CF** over **FA**. Error bars represent one standard deviation above and below the mean.

Metric	Strategy	amh	hau	ibo	lin	lug	orm	pcm	sna	swa	tir	Average Reduction (%)
GPU Memory (GB)	FA	14.5	15.2	15.0	14.7	14.4	14.9	15.3	15.1	14.6	15.0	**33.56**
	CF	**9.8**	**10.1**	**10.0**	**9.9**	**9.7**	**10.0**	**10.2**	**10.0**	**9.8**	**10.1**	
FLOPs (TFLOPs)	FA	21.7	22.8	22.5	22.1	21.6	22.3	23.0	22.7	21.9	22.4	**33.78**
	CF	**14.5**	**14.9**	**14.8**	**14.7**	**14.4**	**14.8**	**15.0**	**14.7**	**14.5**	**14.9**	
Clock Time (Hours)	FA	8.5	9.2	9.0	8.8	8.4	8.9	9.3	9.1	8.6	8.9	**34.83**
	CF	**5.6**	**5.9**	**5.8**	**5.7**	**5.5**	**5.8**	**6.0**	**5.8**	**5.6**	**5.9**	

Table 1: GPU Memory, FLOPs, and Clock Time for MasakhaNEWS dataset using **FA** and **CF**. FLOPs are in TFLOPs, and Clock Time is in hours. Bold values indicate CF's lower computational cost. The last column presents the average percentage reduction of CF compared to FA across all languages.

Metric	Strategy	afr	bem	ewe	fon	ibo	lin	lua	lug	luo	nso	sot	swh	tir	tsn	tso	twi	wol	xho	yor	Average
GPU Memory (GB)	FA	15.2	14.8	14.6	14.9	14.4	14.8	14.6	14.7	14.5	14.9	14.8	15.1	14.8	14.7	14.8	14.6	14.9	14.7	15.0	**31.76**
	CF	**10.1**	**10.0**	**9.9**	**10.0**	**9.7**	**10.0**	**9.9**	**9.8**	**9.9**	**10.1**	**9.9**	**10.2**	**10.1**	**9.9**	**10.0**	**10.1**	**9.9**	**9.8**	**10.0**	
FLOPs (TFLOPs)	FA	22.9	22.5	22.1	22.6	21.8	22.4	22.1	22.2	21.9	22.6	22.3	22.8	22.5	22.0	22.3	21.9	22.7	22.4	23.0	**34.08**
	CF	**14.9**	**14.7**	**14.5**	**14.8**	**14.3**	**14.7**	**14.5**	**14.4**	**14.5**	**14.9**	**14.6**	**15.0**	**14.8**	**14.4**	**14.7**	**14.3**	**14.8**	**14.6**	**15.0**	
Clock Time (Hours)	FA	9.3	9.0	8.7	9.1	8.5	9.0	8.7	8.8	8.6	9.2	8.9	9.3	9.0	8.6	8.8	8.5	9.1	8.9	9.5	**37.08**
	CF	**5.8**	**5.6**	**5.4**	**5.7**	**5.2**	**5.6**	**5.4**	**5.3**	**5.4**	**5.8**	**5.5**	**6.0**	**5.7**	**5.3**	**5.5**	**5.2**	**5.7**	**5.5**	**6.0**	

Table 2: GPU Memory, FLOPs, and Clock Time for SIB-200 dataset using Finetuning All (FA) and Continual Finetuning (CF). FLOPs are in TFLOPs, and Clock Time is in hours. Bold values indicate CF's lower computational cost. The last column presents the average percentage reduction of CF compared to FA across all languages.

plates, their typological distance metrics support our interpretation that CF performs best when languages either appear in pretraining or are typologically close to those that do.

Overall, our analysis reinforces that typological features, particularly language family, script, and morphology, play a central role in the effectiveness of CF. With strong internal cohesion and partial pretraining coverage, Bantu languages benefit more uniformly under CF. In contrast, Afro-Asiatic and West African languages often require more tailored adaptation strategies, and FA provides greater robustness in these cases.

4.3 Finding 3: Uncertainty-Based Selection is Critical for CF Performance

We compare CF with a random acquisition baseline (CF+Random) to isolate the impact of the ac-

Dataset	Statistic	amh	hau	ibo	lug	orm	pcm	run	sna	som	swa	xho	yor
MasakhaNEWS	p-value	0.02	0.07	0.03	0.72	0.03	0.05	0.03	0.02	0.03	0.59	0.03	0.67
	effect size	0.71	3.02	0.35	1.79	0.00	0.38	0.00	0.00	0.00	5.69	0.00	5.00
SIB-200	p-value	-	-	-	0.47	-	-	0.47	-	-	-	0.27	0.07
	effect size	-	-	-	1.34	-	-	1.34	-	-	-	0.89	-

Table 3: Wilcoxon Signed-Rank Test p-values and effect sizes for CF vs. FA across 10 active learning rounds. Each column corresponds to one language. The test compares the F1 scores obtained at each round under CF and FA for each language. For instance, for Amharic (amh), we compute `wilcoxon(cf_scores, fa_scores)`, where each list contains the 10 round-level F1 scores under that setting. A p-value < 0.05 is considered statistically significant. Effect size is computed as $r = \frac{W}{\sqrt{N}}$, where W is the Wilcoxon test statistic and N is the number of paired comparisons.

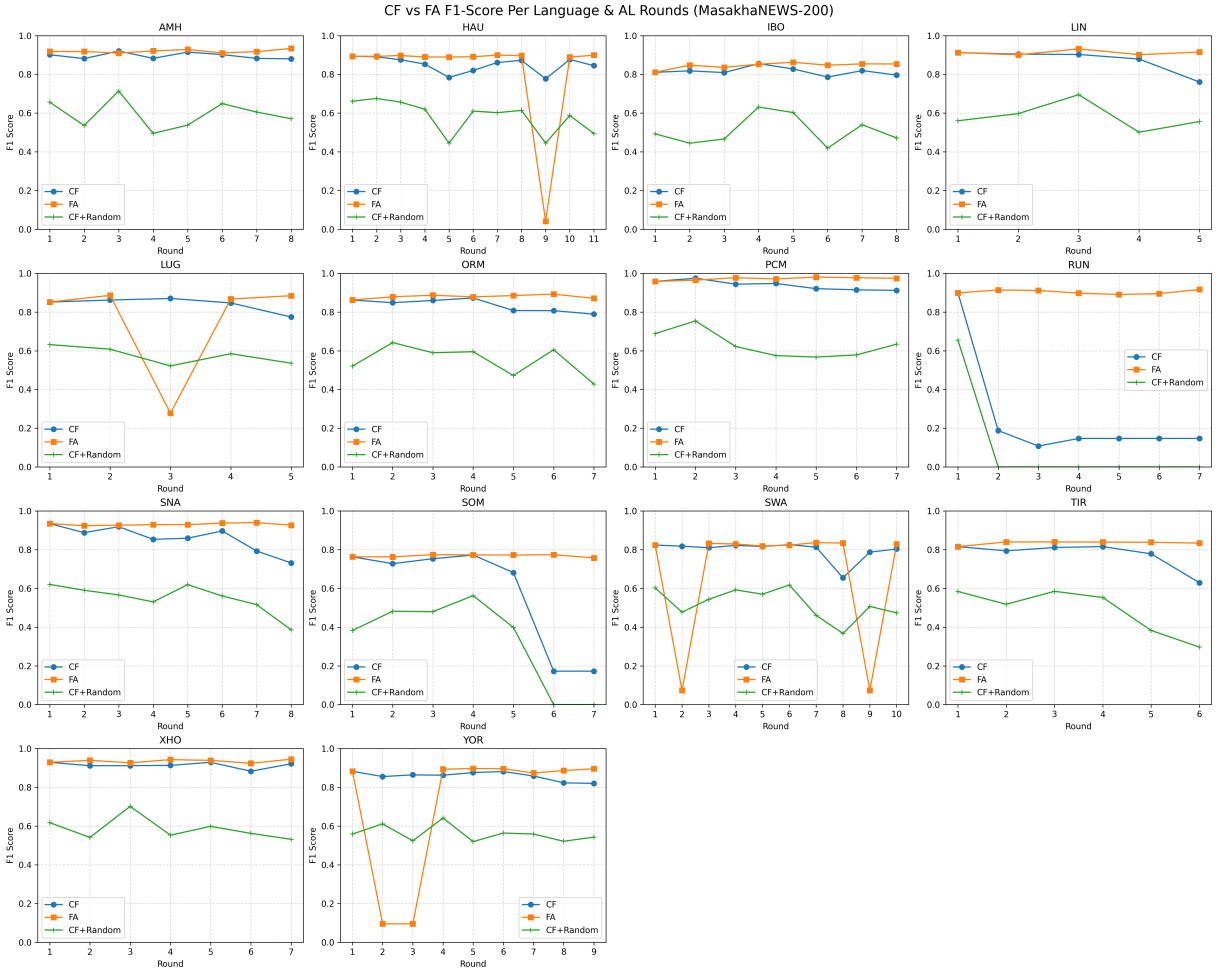

Figure 2: Comparison of CF, FA, and CF+Random across active learning rounds for each language in the MasakhaNEWS-200 dataset. CF consistently matches or closely follows FA, while CF+Random performs significantly worse.

quisition strategy. As shown in Figures 2 and 3, CF+Random underperforms both CF and FA across all languages and rounds. The performance gap is especially pronounced in early and middle rounds, where random selection fails to prioritize informative or uncertain examples.

CF's stateless update mechanism makes it especially reliant on acquiring high-value samples. When guided by uncertainty-based acquisition functions such as Monte Carlo Dropout and entropy scoring, CF receives maximally uncertain and high-gradient inputs, enabling efficient learning. Random acquisition, by contrast, introduces uninformative or redundant samples, which leads to stagnation or regression, particularly in low-resource languages such as Fon (fon), Ewe (ewe),

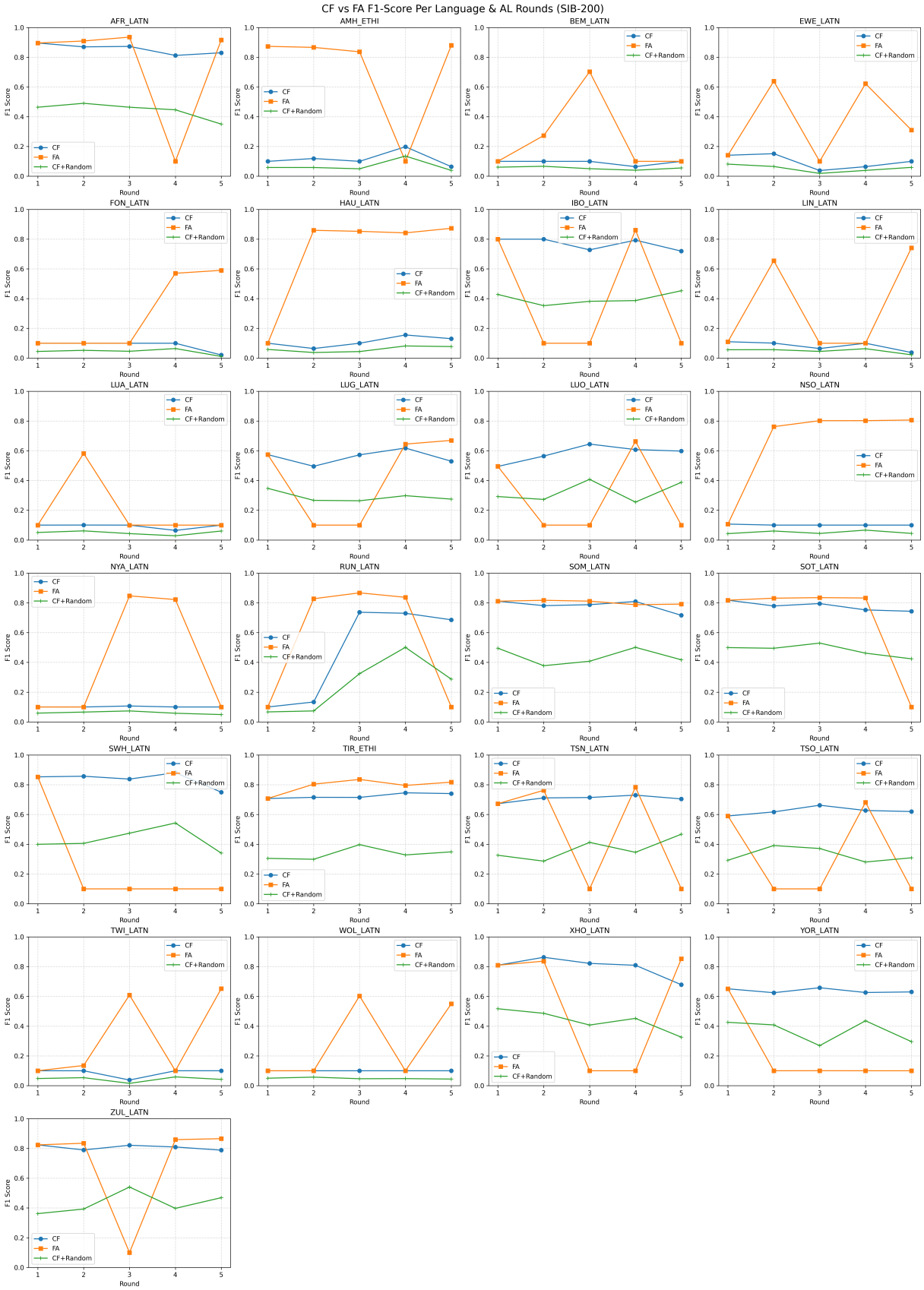

Figure 3: Comparison of CF, FA, and CF+Random across active learning rounds for each language in the SIB-200 dataset. CF maintains comparable performance to FA in most cases, while CF+Random underperforms across the board.

and Tsonga (tso).

Even languages that perform well under CF with uncertainty-based acquisition, like Yoruba (yor) and Xhosa (xho), suffer significant degradation under CF+Random. This confirms that CF's effectiveness depends not only on language similarity or pretraining alignment but also critically on the informativeness of acquired examples.

Moreover, FA, though more stable, is not immune to issues from redundant data. In languages such as Swahili (swa) and Hausa (hau), late-round performance declines under FA, likely due to overfitting to noisy or repetitive samples. These effects are largely mitigated under CF due to its focus on fresh, informative updates.

These findings confirm that uncertainty-based acquisition is helpful and necessary for CF to succeed. In multilingual active learning, the quality of acquired data is often more impactful than quantity.

4.4 Statistical Significance and Dataset-Specific Dynamics

We conducted Wilcoxon signed-rank tests to quantify the consistency of CF versus FA performance across languages. For each language, we collected the F1 scores at each round under CF and FA, respectively, and applied a paired test: `wilcoxon(cf_scores, fa_scores)`. This yielded a p-value assessing whether the per-round scores differ significantly, along with an effect size computed as $r = \frac{W}{\sqrt{N}}$, where W is the Wilcoxon statistic and N is the number of rounds. The results, summarized in Table 3, highlight languages with either statistically significant p-values ($p < 0.05$) or large effect sizes (≥ 0.71).

In MasakhaNEWS, several languages such as Amharic (amh), Igbo (ibo), Oromifa (orm), Runyankore (run), and Shona (sna) show significant p-values, with FA slightly outperforming CF in most of these cases. However, many of these differences are associated with small or even zero effect sizes, indicating limited practical importance. In contrast, languages such as Hausa (hau), Swahili (swa), and Yoruba (yor) display large effect sizes in favor of CF, despite having p-values above the 0.05 threshold. This suggests that CF delivers meaningful but more variable improvements in these cases.

In SIB-200, no languages reach statistical significance. Nevertheless, several languages such as Luganda (lug), Runyankore (run), Xhosa (xho), and Tswana (tsn) exhibit large effect sizes in favor of CF. These results support the broader finding that CF performs particularly well in controlled, low-resource environments with consistent acquisition conditions.

These trends are driven by the structural differences between the two datasets. MasakhaNEWS contains languages with highly variable training sizes, ranging from 608 examples for Lingala (lin) to over 3,300 for English (eng), as well as unbalanced label distributions. These characteristics increase the likelihood of overfitting under FA, especially in later rounds. In contrast, SIB-200 follows a uniform structure with around 1,000 samples per language and balanced splits. This setup favors the stateless nature of CF by providing consistent learning signals across rounds.

These findings confirm that CF is an effective option in stable, multilingual settings, offering significant computational savings without major loss in accuracy. FA may still be necessary for languages with weaker pretraining alignment, unstable learning dynamics, or pronounced data imbalance. Future research should explore adaptive finetuning strategies that dynamically select CF and FA based on acquisition quality, statistical variance, or round-level learning signals.

5 Conclusion

This work re-examines the assumption that FA is necessary in AL, especially for African languages with limited data and computational resources. We evaluate *Continual Finetuning (CF)* as a resource-efficient alternative and find that it substantially reduces computational resources, while delivering performance comparable to (*FA*) in most settings. (1) CF performs best when the target language is included in the model's pretraining corpus, where strong initialization and adequate supervision lead to stable learning dynamics. (2) CF can also be effective for non-pretraining languages that are typologically close to pretraining ones, particularly Bantu languages, thanks to shared linguistic structures. (3), CF's success depends critically on uncertainty-based acquisition; without it, performance degrades sharply, highlighting the need for principled sample selection. Although FA still outperforms CF in some instances, particularly for languages with unstable acquisition dynamics, limited pretraining overlap, or high label imbalance, these gains often come with modest effect sizes. Overall, CF emerges as a strong alternative for low-resource multilingual AL pipelines, and these

findings motivate the development of hybrid strategies that adaptively switch between CF and FA based on acquisition signals, typological features, or confidence variance. Our study builds scalable, inclusive, and efficient learning systems for underrepresented languages.

6 Broader Impacts

This work explores active learning strategies for improving NLP models for African languages. By enabling more efficient and cost-effective model training, particularly in low-resource settings, our approach can help close the performance gap for underrepresented languages. This supports linguistic equity and inclusivity efforts in AI technologies, especially in regions with limited computational resources and access to annotated data.

Positive Impacts: Our method reduces the need for extensive computational resources and large-scale annotated datasets. This democratizes access to language technologies by allowing researchers and practitioners in low-resource settings to build useful models with fewer resources. Moreover, by enhancing the performance of African language models, this work can contribute to more equitable digital access, promote civic participation, and support educational, governmental, and cultural initiatives within African communities.

Potential Negative Impacts: As with any technology that enables easier deployment of NLP models, there is a risk of misuse, such as deploying under-tested systems in sensitive applications (e.g., health, law, or government) without proper safeguards or validation. Additionally, more efficient model training may inadvertently promote the development of systems without community involvement, potentially reinforcing language representation biases if datasets are not carefully curated.

We encourage future work to include affected communities in the design, deployment, and evaluation processes. Fair and transparent data practices remain essential to ensure that efficiency gains do not come at the cost of ethical responsibility.

7 Limitations

While continual finetuning significantly reduces computational costs, it may lead to performance degradation for languages not seen during pretraining. Full finetuning remains more stable in such cases, suggesting that continual finetuning alone may not be optimal for all settings. Future work could explore adaptive strategies that selectively apply full finetuning when performance instability is detected, balancing efficiency and effectiveness across different language scenarios.

Acknowledgements

We thank the Mila computational resources used in this study. We also acknowledge funding support from the Canada CIFAR AI Chair program.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023a. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *Preprint*, arXiv:2309.07445.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023b. MasakhaNEWS: News topic classification for African languages. In *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 (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

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 Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumim, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Divyanshu Aggarwal, Sankarshan Damle, Navin Goyal, Satya Lokam, and Sunayana Sitaram. 2024. Towards exploring continual fine-tuning for enhancing language ability in large language model. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ali Ayub and Carter Fendley. 2022. Few-shot continual active learning by a robot. In *Advances in Neural Information Processing Systems*.

Javad Azimi, Alan Fern, Xiaoli Zhang-Fern, Glencora Borradaile, and Brent Heeringa. 2012. Batch active learning via coordinated matching. *arXiv preprint arXiv:1206.6458*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

Arnav Mohanty Das, Gantavya Bhatt, Megh Manoj Bhalerao, Vianne R. Gao, Rui Yang, and Jeff Bilmes. 2023. Accelerating batch active learning using continual learning techniques. *Transactions on Machine Learning Research*.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bonaventure FP Dossou. 2023. Advancing african-accented speech recognition: Epistemic uncertainty-driven data selection for generalizable asr models. *arXiv preprint arXiv:2306.02105*.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017a. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017b. Deep bayesian active learning with image

data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1183–1192. JMLR.org.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng

Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuhong Guo and Dale Schuurmans. 2008. Discriminative batch mode active learning. In *Advances in neural information processing systems*, pages 593–600.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, George van den Driessche, Aurelia Guy, Brooks Paige, Phil Withers, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. 2006. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424. ACM.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Jonas Hübotter, Lenart Treven, Yarden As, and Andreas Krause. 2024. Transductive active learning: Theory and applications. *Advances in Neural Information Processing Systems*, 37:124686–124755.

Moksh Jain, Emmanuel Bengio, Alex-Hernandez Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. 2023. Biological sequence design with gflownets. *Preprint*, arXiv:2203.04115.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. *BatchBALD: efficient and diverse batch acquisition for deep Bayesian active learning*. Curran Associates Inc., Red Hook, NY, USA.

Andreas Kirsch, Sebastian Farquhar, Parmida Atighehchian, Andrew Jesson, Frederic Branchaud-Charron, and Yarin Gal. 2023. Stochastic batch acquisition: A simple baseline for deep active learning. *Preprint*, arXiv:2106.12059.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12.

Seiji Maekawa, Dan Zhang, Hannah Kim, Sajjadur Rahman, and Estevam Hruschka. 2022. Low-resource interactive active labeling for fine-tuning language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3230–3242, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Martin Mundt, Yongwon Hong, Iuliia Pliushch, and Visvanathan Ramesh. 2023. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *Neural Netw.*, 160(C):306–336.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi,

Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Matthieu Texier, and Jeffrey Dean. 2021. The carbon footprint of machine learning training will plateau, then shrink. *arXiv preprint arXiv:2104.10350*.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Model Training Hyperparameter	Value
Model Name	Davlan/afro-xlmr-large
Evaluation Strategy	steps
Save Strategy	steps
Save Steps	50000
Learning Rate	5e-5
Per Device Train Batch Size	16
Per Device Eval Batch Size	16
Num Train Epochs	10
Weight Decay	0.01
Logging Steps	10000
Save Total Limit	1
Load Best Model at End	True
Max Length	128
Active Learning Sample Selection	
Pool Size	0.5 (50% of training set)
Number of MC Dropout Passes	10
Top-K Uncertainty Samples	100

Table 4: Hyperparameters used for model and sample selection in the active learning loop.

Dataset	Languages
MasakhaNEWS	Amharic (amh) Hausa (hau) Igbo (ibo) Lingala (lin) Luganda (lug) Oromo (orm) Nigerian Pidgin (pcm) Kirundi (run) Shona (sna) Somali (som) Swahili (swa) Tigrinya (tir) Xhosa (xho) Yoruba (yor)
SIB-200	Amharic (amh_Ethi) Afrikaans (afr_Latn) Bemba (bem_Latn) Ewe (ewe_Latn) Fon (fon_Latn) Hausa (hau_Latn) Igbo (ibo_Latn) Lingala (lin_Latn) Luba-Kasai (lua_Latn) Luo (luo_Latn) Luganda (lug_Latn) Northern Sotho (nso_Latn) Nyanja (nya_Latn) Kirundi (run_Latn) Somali (som_Latn) Sotho (sot_Latn) Swahili (swh_Latn) Tswana (tsn_Latn) Tigrinya (tir_Ethi) Tsonga (tso_Latn) Twi (twi_Latn) Wolof (wol_Latn) Xhosa (xho_Latn) Yoruba (yor_Latn) Zulu (zul_Latn)

Table 5: Languages used in the MasakhaNEWS and SIB-200 datasets.