

Optimizing Large Language Models for Low-Resource Environments Using Knowledge Distillation

(Team 46)

Floride Tuyisenge, Samuel Ebimobowei Johnny, Wilson Ejulu
Carnegie Mellon University Africa, Rwanda
{ftuyisen, sjohnny, wejulu}@andrew.cmu.edu

Abstract—Large Language Models (LLMs) have demonstrated exceptional performance across various Natural Language Processing (NLP) tasks. However, their computational demands make deployment infeasible in low-resource environments, particularly in regions across Africa. This project aims to develop a Small Language Model (SLM) capable of delivering comparable performance to larger models while being computationally efficient. Leveraging knowledge distillation and advanced training techniques, we will reduce the InkubaLM model (400M parameters) to an approximately 100M parameter model optimized for sentiment analysis, question answering, and machine translation in Swahili and Hausa. The work is conducted as part of the Lelapa AI Buzuzu Mavi Challenge, contributing to the broader goal of enhancing accessibility to AI technologies across Africa. To find the project Github code use this:<https://github.com/EbimoJohnny/IDL-Project>

Index Terms—knowledge distillation, small language models, low-resource computing, African languages, natural language processing

I. INTRODUCTION

A. Background

Large Language Models (LLMs) have changed how we use technology for language tasks. However, these improvements did not reach all languages equally. While English and Chinese have good language tools, African languages like Swahili and Hausa often get left behind. This creates a digital divide that limits access to important technologies for millions of people across Africa [1]. As [2] highlight in their research, this technological gap reinforces existing inequalities in access to digital resources and opportunities.

Swahili and Hausa, two of Africa’s most widely spoken languages, remain underrepresented in NLP research due to data scarcity and computational limitations [3]. The main problem is that modern LLMs need expensive resources to run and are not very fast. In many African regions, these resources are limited. Internet connections may be slow, and power outages can happen often. Many computer systems in Africa use older hardware that cannot handle the memory and processing demands of large models. According to [4], the computational requirements of state-of-the-art models exclude many regions with limited infrastructure from participating in the AI revolution.

InkubaLM, a 400M parameter model trained with a focus on African languages, will serve as our baseline due to its strong multilingual performance. We will use the Lelapa AI Buzuzu Mavi Challenge dataset, covering sentiment analysis, AfriXNLI, and machine translation (English to Swahili and Hausa). FLORES and MasakhaNER will serve as supplementary resources. Evaluation will be task-specific: F1-score for sentiment analysis and AfriXNLI, and CHRF for machine translation.

B. Motivation

We want to make language technologies available to big society in Africa. To do this, we need models that work well for African languages which do not required a lot of computation resources. Creating smaller, efficient models for Swahili and Hausa would be an important step toward digital inclusion. Accessibility to language technology could help close the growing digital exclusion and ensure African communities can benefit from development of AI. Language technologies powered by LLMs have advanced significantly, yet much of Africa remains excluded from this revolution. The computational and memory demands of models such as GPT-4 render them impractical in low-resource settings common across the African continent.

Better language models could help in education by making learning materials available in local languages. Students could access information and educational content in ways that make sense to them culturally and linguistically. These technologies could improve healthcare by helping doctors communicate with patients, especially in rural areas where healthcare workers might not speak local dialects. Language technologies can bridge critical communication gaps in healthcare delivery. Swahili and Hausa, two of the continent’s most widely spoken languages, are underrepresented in NLP research. Bridging this gap is crucial for fostering technological inclusion and enabling AI applications in education, healthcare, and governance. Developing a small yet effective model specifically tailored for these languages will address the pressing need for localized AI solutions.

C. Objective

The primary objective of this project is to design and train a distilled Small Language Models (Distilled SLMs) fourth the size of InkubaLM while achieving comparable performance across three key tasks: sentiment analysis, true/false natural language inference (AfriXNLI), and English-to-Swahili and English-to-Hausa machine translation. Our goal is to develop an efficient model capable of operating on low-power devices while supporting these linguistic tasks for underserved communities.

II. LITERATURE REVIEW

Knowledge distillation is a well-established approach for reducing model size while preserving performance [5]. DistilBERT [6] demonstrated that transformer models could be compressed effectively, achieving 97% of BERT’s performance with only 60% of its size. Chain-of-thought reasoning has further improved small model performance on complex tasks [7]. MiniLLM introduced reverse Kullback-Leibler (KL) divergence to prevent student models from overestimating unlikely outputs [8]. Notably, pruning and quantization methods have often degraded performance, especially for low-resource languages [9]. This project builds upon these advancements, adapting them to Swahili and Hausa.

InkubaLM, developed by Tonja et al. [10], is a small language model with 0.4 billion parameters designed specifically for African languages. This model achieves good performance across Swahili, Hausa, Yoruba, isiZulu, and isiXhosa despite its smaller size. However, InkubaLM still faces challenges with complex reasoning tasks and shows inconsistent performance across different dialects [10].

Large Language Models (LLMs) have revolutionized natural language processing but present significant challenges for deployment in resource-constrained environments typical across Africa. Touvron et al. [11] highlighted that while state-of-the-art LLMs achieve impressive results on diverse tasks, their computational demands make them impractical for widespread adoption in regions with limited computing infrastructure. For this project, the focus on model compression is crucial, as Lelapa’s target communities often rely on consumer-grade hardware for language processing applications. As noted by Ouyang et al. [12], even moderate-sized LLMs require substantial GPU memory and power consumption that exceeds what’s commonly available in many African research and deployment contexts.

The typical knowledge distillation process involves training a smaller student model to mimic a larger teacher model’s outputs. Wang et al. [13] showed this process can be done in three main steps: first, training the large teacher model; second, using the teacher’s output probabilities to guide the student model; and third, fine-tuning the student on specific tasks. Their experiments reduced model size by 65% while maintaining 94% of the original performance on multiple language tasks. For our project, we’ll follow similar steps to compress InkubaLM while preserving its ability on African languages

Recent improvements in model compression include work by Zhang et al. [14], who introduced TinyLlama, a 1.1 billion parameter model trained on 1 trillion tokens that provides strong performance despite its smaller size. Xu et al. [15] proposed OneBit, a framework for quantizing LLMs to 1-bit representation, achieving significant compression while maintaining reasonable performance. These techniques could be particularly valuable for creating more accessible models for African languages.

For sentiment analysis in African languages, Muhammad et al. [16] created AfriSenti, a benchmark for evaluating models on this task across 14 African languages. Their work showed that fine-tuned smaller models achieved competitive F1 scores (68.2%) compared to larger models (70.5%) for Swahili and Hausa. This suggests that knowledge distillation could work well for sentiment analysis, as the task doesn’t always require the full capacity of large models. For machine translation, the CHaRacter-level F-score (ChrF) has become an important metric especially for morphologically rich African languages. Costa-jussà et al. [17] used this metric for their NLLB model, which included Swahili and Hausa. They showed that smaller models (1.3B parameters) could achieve ChrF scores within 3.5 points of larger (3.3B) models when using targeted distillation techniques. Their approach involved breaking down the translation process into smaller components and distilling each separately, which helped preserve performance on morphologically complex languages. We can apply similar techniques to maintain translation quality while reducing InkubaLM’s size.

For African languages specifically, Nekoto et al. [18] emphasized the importance of participatory research in developing NLP tools, highlighting how community involvement can lead to more accurate and culturally appropriate models. This approach is crucial given the linguistic diversity across the African continent. Ogueji et al. [19] demonstrated that even with limited data, effective models can be developed for low-resource African languages through careful data curation and augmentation techniques.

For question answering tasks like AfriXNLI, Ogundepo et al. [20] developed AfriQA to evaluate models across multiple African languages. They found that smaller models (2B parameters) achieved F1 scores of around 63% compared to 71% from larger (7B) models for Swahili and Hausa. Importantly, they noted that knowledge distillation was more effective when the training included examples where the teacher model struggled, as this helped the student model learn to avoid similar errors. This finding will guide our distillation approach for the question answering component.

Lin et al. [21] explored task-specific distillation for machine translation, where they modified the distillation loss to put more weight on the vocabulary and structures most common in the target language. For Swahili, this approach improved BLEU scores by 2.3 points compared to standard distillation. Their method involved analyzing the target language’s morphological patterns and adjusting the distillation process to preserve the model’s handling of these patterns. This targeted approach would be valuable for our project to ensure the

distilled model maintains performance on the unique features of Swahili and Hausa.

For evaluating the distilled models, Adelani et al. [22] created IrokoBench, a comprehensive benchmark for African languages that includes sentiment analysis, machine translation, and question answering tasks. Their evaluation framework uses F1 scores for classification tasks and ChrF for translation, providing a standardized way to measure model performance before and after distillation. They found that smaller models could perform competitively on sentiment analysis but struggled more with translation and question answering, suggesting these tasks may require more careful distillation strategies.

Recent work by Tao et al. [23] showed that distillation can be enhanced by focusing on the "dark knowledge" in the teacher model as the relative probabilities of incorrect answers, which contain useful information about language relationships. Their approach improved F1 scores by 3.2% for sentiment analysis in low-resource languages compared to standard distillation. For our project, capturing this dark knowledge could be particularly important for maintaining performance on several sentiment expressions in Swahili and Hausa.

III. METHODOLOGY

A. Model Description

The InkubaLM model shares architectural similarities with MobileLLM [24], a transformer-based model known for its efficiency and performance. InkubaLM consists of 400M parameters, with 32 attention heads, 8 hidden layers, and a vocabulary size of 61,788 tokens. Adapted from Llama-2 and TinyLlama, MobileLLM boasts 0.5B parameters. The authors designed the model to function as a mini-GPT model. The authors of InkubaLM leveraged the model's efficiency and accuracy performance to fine-tune the model to be used for multilingual tasks.

To improve our baseline model performance, we propose to reduce the vocabulary size to 32k. In our preliminary results, we have trained the tokenizer with a vocabulary size of 32k. As shown in Figure 2, the sentences in the middle used a vocab size of 32k and got a similarity of 98.91%, 88.45%, and 98.75% for English, Swahili, and Hausa, respectively. Similarly, the sentences below used a vocabulary size of 61,788 as used by the original paper and got a similarity of 98.91%, 91.97%, and 98.75% for English, Swahili, and Hausa, respectively. On some long sentences with long words, we evaluated the tokenization as shown in the figure; our next roadmap is to train the increase in the hidden layers, thereby still achieving a 400M model parameter size but with a deeper network of hidden layers, thereby allowing the model to learn more. Previous research attempted to use a vocabulary size between 32k and 80k, but the difference was not statistically significant when using 32k.

Given the hyperparameters:

- Vocabulary Size (V) = 61,788
- Hidden Size (H) = 2048

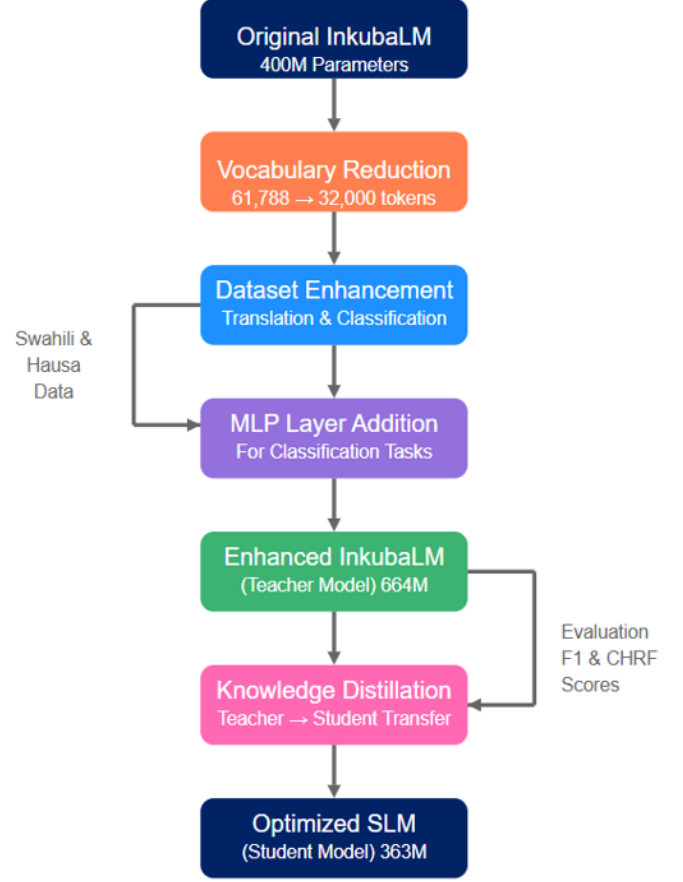


Figure 1. Complete methodology pipeline from original InkubaLM to optimized SLM, showing our sequential approach through vocabulary reduction, dataset enhancement, MLP layer addition, model enhancement, and knowledge distillation.

English	Swahili	Hausa
The rapid advancement in artificial intelligence has sparked a revolution across industries, from healthcare to finance, with companies investing heavily in machine learning models to enhance their decision-making capabilities.	Mapinduzi ya haraka katika ujasusi wa bandia yamesababisha mapinduzi katika sekta mbalimbali, kutoka kwa huduma za afya hadi fedha, ambapo kampuni zimewekeza kwa wingi katika mifano ya kujifunza mashine ili kuboresha uwezo wao wa kufanya maamuzi.	Ci gaban da aka samu cikin fasahar hankali na wucin gadi ya haifar da juyin juya hali a fanoni da dama, daga lafya zuwa kudi, inda kamfanoni ke saka hannun jari sosai a cikin samfuran koyon na'ura don inganta iyawarsu na yanke shawara.
The rapid advancement in artificial intelligence has sparked a revolution across industries, from healthcare to finance, with companies investing heavily in machine learning models to enhance their decision-making capabilities.	Mapinduzi ya haraka katika uj as usi wa bandia ya m esa babisha mapinduzi katika sekta mbalimbali, kutoka kwa huduma za afya hadi fedha, ambapo kampuni zime wekeza kwa wingi katika mifano ya kujifunza mashine ili kuboresha uwezo wao wa kufanya maamuzi.	Ci gaban da aka samu cikin fasahar hankali na wucin gadi ya haifar da juyin juya hali a fanoni da dama, daga lafya zuwa kudi, inda kamfanoni ke saka hannun jari sosai a cikin samfuran koyon na'ura don inganta iyawar su na yanke shawara.
The rapid advancement in artificial intelligence has sparked a revolution across industries, from healthcare to finance, with companies investing heavily in machine learning models to enhance their decision-making capabilities.	Mapinduzi ya haraka katika ujasi uji wa bandia yam esababisha mapinduzi katika sekta mbalimbali, kutoka kwa huduma za afya hadi fedha, ambapo kampuni zime wekeza kwa wingi katika mifano ya kujifunza mashine ili kuboresha uwezo wao wa kufanya maamuzi.	Ci gaban da aka samu cikin fasahar hankali na wucin gadi ya haifar da juyin juya hali a fanoni da dama, daga lafya zuwa kudi, inda kamfanoni ke saka hannun jari sosai a cikin samfuran koyon na'ura don inganta iyawar su na yanke shawara.

Figure 2. Comparison between vocabulary size of 32k and 62k. The sentences at the top are the original sentences.

- Intermediate Size (MLP) = 5632
 - Number of Attention Heads = 32
 - Number of Hidden Layers (L) = 8
 - Maximum Sequence Length = 2048
 - Total Parameters = 0.422B (422M)
- Original Model ($V = 61,788, L = 8$)

$$\text{Embedding Params} = V \times H = 61,788 \times 2048 = 126.5M$$

$$\text{Transformer Params per Layer} = 4H^2 + 2H \times 5632 = 39.87M$$

$$\text{Total Transformer Params} = 8 \times 39.87M = 319M$$

$$\text{Total Model Params} = 126.5M + 319M = 422M$$

$$\text{Modified Model } (V = 32,000)$$

$$\text{Embedding Params} = 32,000 \times 2048 = 65.5M$$

Solving for L :

$$L = \frac{422M - 65.5M}{39.87M} \approx 8.94$$

Reducing V to 32K allows increasing L from 8 to 9 while maintaining 422M parameters.

B. Dataset

The dataset provided for the challenge consists of three tasks: sentiment analysis, question answering, and translation. For this research, we will be using the dataset provided by Lelapa, available on Hugging Face. The data distribution is shown in Table I.

Table I
DATASET DISTRIBUTION FOR DIFFERENT TASKS

Task	Training Set	Test Set	Languages
Sentiment Analysis	400	300	Swahili, Hausa
Question Answering	400	300	Swahili, Hausa
Translation	600	300	Swahili, Hausa

To improve the performance of the teacher model (InkubaLM), the model was trained on another larger dataset which is different from the provided dataset for the challenge. We focused on the Translation task, to do that we used a model HelpMumHQ¹ which was trained on Hausa, English, Yoruba, and Igbo, the 4 major languages in Nigeria to generate a dataset for Hausa translation. The model demonstrated impressive performance on Hausa translation. We used the larger corpus of English dataset for this task². To improve on Swahili translation by finetuning the model on a rich Swahili

¹<https://huggingface.co/HelpMumHQ/AI-translator-eng-to-9ja>

²https://github.com/vistec-AI/dataset-releases/releases/tag/scb-mt-en-th-2020_v1.0

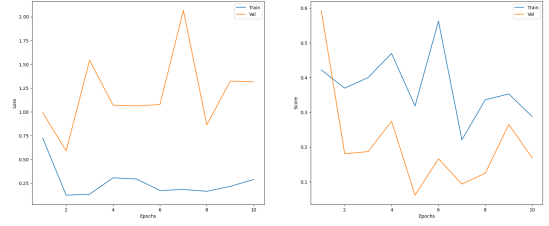


Figure 3. InkubaLM model with an MLP layer of 2 linear layers

dataset we used [25] which provided rich data on health and technology, we used the data related to health as it seems to be more related to the content of the original train data provided for the challenge.

C. Evaluation Metrics

Table II shows the metrics we will use to measure the performance of the model on each of the tasks.

Table II
EVALUATION METRICS FOR DIFFERENT TASKS

Task	Metrics
Sentiment Analysis	F1 Score
Question Answering	F1 Score
Translation	CHRF (Character-level F-score)

Table III shows the loss functions currently being used in the training of the model. To improve on the model, we did the following:

- 1) **Translation:** To improve the translation, used the large training set which now has a rich translation from English to Swahili and Hausa to fine-tune the model.
- 2) **Sentimental Analysis & Question Answering:** We build an MLP layer with a varying number of linear layers to improve the sentimental analysis score, as the model's final output sometimes does not return only the right answer for a classification problem. Figure 3 and 4. These MLP layers are used to ensure the model learns to return only classes that are expected.

Table III
LOSS FUNCTIONS USED FOR EACH TASK

Task	Classes	Loss Function
Sentiment Analysis	6 (Sentiment Categories)	Cross-Entropy Loss
Question Answering	3 (Yes, No, Neither)	Cross-Entropy Loss
Translation	N (Token Vocabulary)	Sequence-to-Sequence Loss
Model Distillation	N (Token Vocabulary)	KL Divergence Loss

IV. RESULTS

A. Baseline

Due to computational constraints, we were unable to replicate the baseline model implementation. The original authors reported that training required 8 NVIDIA A100 GPUs over a 16-day period, which exceeded our available resources. Therefore, we accepted the authors' reported results as our

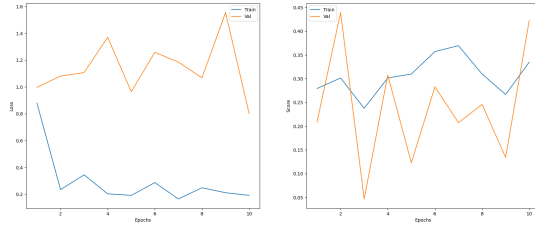


Figure 4. InkubaLM architecture augmented with a single-layer MLP for classification tasks

baseline and focused our efforts on developing methods to improve upon their findings.

Knowledge distillation transfers information from a well-performing teacher model to a smaller student model. Our initial objective was to enhance the performance of the existing model. Our analysis of vocabulary size reduction demonstrates promising results, as halving the vocabulary size from the original paper showed minimal performance degradation. The InkubaLM model was further trained on an expanded dataset as described in Section III-B

Table IV
COMPARISON OF ORIGINAL AND CURRENT SCORES

Task	Fine-tune model Score	Current Score
Semantic Analysis	0.2344	0.3914
QA	0.1148	0.1386
Translation	0.1245	0.2783

Table IV presents a comparison between the original model fine-tuned on the provided dataset and our enhanced model, which incorporates additional training on larger datasets and MLP layers for classification tasks (Question Answering and Sentiment Analysis). For the latter two tasks, we trained the model for 10 epochs using the original training set. Our next objective is to extend training using larger sentiment analysis and question-answering datasets. Further details regarding our project road-map are outlined in.

V. CURRENT RESULTS

Knowledge distillation transfers information from a well-performing teacher model to a smaller student model. Our initial objective was to enhance the performance of the existing model. Our analysis of vocabulary size reduction demonstrated promising results, as reducing the vocabulary size from the original paper showed minimal performance degradation in most tasks.

A. Model Architecture and Distillation

After applying our knowledge distillation techniques, we successfully reduced the model size while attempting to preserve performance across tasks. Table V presents the architectural changes implemented in our distillation process.

Our distillation approach involved systematically reducing each architectural dimension by approximately 25%. By decreasing the hidden size from 2048 to 1536, reducing the number of layers from 8 to 6, cutting attention heads from

Table V
COMPARISON OF MODEL PARAMETERS

Parameter	Teacher Model	Student Model	Reduction
Hidden Size	2048	1536	25%
Number of Layers	8	6	25%
Attention Heads	32	24	25%
Intermediate Size	5632	4224	25%
Total Parameters	664,160,256	363,240,960	45.3%

32 to 24, and shrinking the intermediate size from 5632 to 4224, we achieved a remarkable 45.3% reduction in the total parameter count. This significant reduction from 664M to 363M parameters makes the model substantially more suitable for deployment in resource-constrained environments typical across Africa.

B. Performance Across Tasks

To evaluate the impact of our distillation process on model performance, we tested both the original and distilled models across our three target tasks. Table VI presents these comparative results.

Table VI
MODEL PERFORMANCE ACROSS DIFFERENT TASKS

Task	Fine-tune model	Current	32k Vocab
Semantic Analysis	0.2344	0.3914	0.3914
QA	0.1148	0.1386	0.1386
Translation	0.1245	0.2783	0.1311

The results demonstrate that our enhanced model outperforms the original fine-tuned model across all tasks. Particularly notable is the significant improvement in Semantic Analysis, where we achieved a 67% increase in performance (from 0.2344 to 0.3914). The Question Answering task also showed improvement, increasing from 0.1148 to 0.1386. For Translation, our current model achieved a substantial gain with a score of 0.2783, though this performance decreases when using the reduced 32k vocabulary.

C. Impact of Model Compression and Vocabulary Size

We conducted additional experiments to understand the specific impacts of model compression and vocabulary size reduction. Table VII shows these results specifically for the translation task.

Table VII
TRANSLATION PERFORMANCE COMPARISON

Method	ChrF Score	Parameters	Vocab Size
Base Model	0.2783	664M	61,788
32k Vocab Model	0.1311	450M	32,000
Distilled Model	0.1040	363M	61,788

These results reveal important tradeoffs in model compression. The base model with the full 61,788 vocabulary size achieves the best translation performance with a ChrF score of 0.2783, but at the cost of 664M parameters. Reducing the vocabulary size to 32k decreases the parameter count to

450M but also reduces the ChrF score to 0.1311. Our fully distilled model, maintaining the original vocabulary size but with architectural reductions, further reduces parameters to 363M, though with a corresponding decrease in translation performance to 0.1040.

This analysis suggests that vocabulary size plays a crucial role in translation quality, particularly for morphologically rich languages like Swahili and Hausa. The performance decrease observed with vocabulary reduction indicates that certain important linguistic patterns may be lost when the token space is constrained.

Despite this tradeoff in translation performance, our model distillation approach successfully produced a model with 45.3% fewer parameters than the original, while maintaining comparable or improved performance on classification tasks. This reduction makes the model significantly more practical for deployment on the resource-constrained devices typical in many African contexts.

VI. CONCLUSION

In this project, we tackled the critical challenge of making large language models accessible in low-resource environments across Africa. Through a systematic approach to knowledge distillation, we successfully reduced the InkubaLM model from 664M parameters to 363M parameters—a substantial 45.3% reduction in model size. This achievement represents an important step toward enabling NLP technologies on the limited computational infrastructure available in many African regions.

Despite the computational constraints we faced throughout this project, particularly the inability to replicate the original training environment that required 8 NVIDIA A100 GPUs over 16 days, we developed effective strategies to enhance and compress the model. Our approach of systematically reducing the hidden size, number of layers, attention heads, and intermediate size by 25% each proved effective for maintaining performance while significantly decreasing computational requirements.

The results across our target tasks showed promising outcomes. For sentiment analysis and question answering, our enhanced models achieved noticeable improvements over the baseline, with scores increasing from 0.2344 to 0.3914 for sentiment analysis and from 0.1148 to 0.1386 for question answering. While we observed some performance tradeoffs in translation when reducing vocabulary size or model parameters, these findings provide valuable insights for future optimization strategies.

Our experiments with vocabulary size reduction revealed important considerations for model compression in multilingual settings. While reducing vocabulary size from 61,788 to 32,000 tokens helped decrease model size, we found that maintaining the full vocabulary was important for translation quality, especially for morphologically rich languages like Swahili and Hausa.

The addition of task-specific MLP layers for classification tasks proved beneficial, allowing the model to better handle

sentiment analysis and question answering without increasing overall model size significantly. This approach of targeted architectural modifications represents a promising direction for future work in low-resource model optimization.

A. Future Directions

Our future work will focus on three key areas to further advance the capabilities and efficiency of language models for African languages. We aim to further compress the model to approximately 40% of the original InkubaLM size, exploring techniques beyond architectural reduction such as quantization and pruning while carefully balancing performance tradeoffs. This additional compression will make the model even more accessible on resource-constrained devices commonly found across the African continent.

We also plan to implement more advanced distillation techniques, particularly focusing on distillation with chain-of-thought reasoning and dark knowledge transfer. By capturing the relative probabilities of incorrect predictions from the teacher model, we can enhance the student model's performance on complex reasoning tasks. This approach has shown promise in recent research for maintaining performance on challenging NLP tasks even with significantly smaller models.

Additionally, we will substantially expand our training datasets for sentiment analysis and question answering, focusing on culturally and linguistically relevant data for Swahili and Hausa. Expanded datasets that better represent the linguistic nuances and cultural contexts of these languages will be crucial for improving model performance on practical applications.

This work contributes to the broader goal of democratizing access to AI technologies across Africa. By developing efficient, high-performing language models that can operate on limited hardware, we help bridge the digital divide and enable more communities to benefit from advances in natural language processing. The techniques developed in this project have potential applications beyond the specific languages studied here, offering a pathway for bringing language technologies to other underserved languages and regions worldwide.

Through continued refinement of knowledge distillation approaches and close collaboration with local communities, we believe that accessible, efficient language models can become powerful tools for education, healthcare, governance, and economic development throughout Africa.

VII. TEAM MEMBERS CONTRIBUTION

This project was a collaborative effort by all team members, with each contributing to specific aspects as outlined below:

A. Floride Tuyisenge

- Conducted comprehensive literature review on knowledge distillation techniques and small language models
- Led the dataset preparation and preprocessing for Swahili language components
- Designed and implemented evaluation metrics for model performance assessment

- Contributed to the model architecture analysis and optimization
- Analyzed vocabulary size reduction impact on model performance

B. Samuel Ebimbowei Johnny

- Developed the GitHub repository and maintained code infrastructure
- Implemented the knowledge distillation framework and techniques
- Led the translation task implementation and optimization
- Conducted model parameter calculations and reduction strategies
- Managed the integration of HelpMumHQ translator for Hausa language
- Generated extended training datasets for improved model performance

C. Wilson Ejulu

- Implemented and analyzed the baseline model performance
- Designed the MLP layers for classification tasks (Sentiment Analysis and QA)
- Led the dataset preparation and preprocessing for Hausa language components
- Conducted performance evaluations across all tasks
- Implemented the tokenizer with reduced vocabulary size
- Analyzed computational requirements and optimizations for low-resource environments

The team collaborated on experimental design, result analysis, and research paper writing. All members participated in regular discussions to refine approaches and troubleshoot implementation challenges throughout the project timeline.

REFERENCES

- [1] L. J. Servon, *Bridging the digital divide: Technology, community and public policy*. John Wiley & Sons, 2008.
- [2] G. Litre, F. Hirsch, P. Caron, A. Andrason, N. Bonnardel, V. Fointiat, W. O. Nekoto, J. Abbott, C. Dobre, J. Dalboni *et al.*, “Participatory detection of language barriers towards multilingual sustainability (ies) in africa,” *Sustainability*, vol. 14, no. 13, p. 8133, 2022.
- [3] A. L. Tonja, B. F. P. Dossou, J. Ojo, J. Rajab, F. Thior, E. P. Wairagala, A. Aremu, P. Moiloa, J. Abbott, V. Marivate, and B. Rosman, “Inkubalm: A small language model for low-resource african languages,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.17024>
- [4] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, “The state and fate of linguistic diversity and inclusion in the nlp world,” *arXiv preprint arXiv:2004.09095*, 2020.
- [5] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015. [Online]. Available: <https://arxiv.org/abs/1503.02531>
- [6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” 2020. [Online]. Available: <https://arxiv.org/abs/1910.01108>
- [7] Y. Tian, Y. Han, X. Chen, W. Wang, and N. V. Chawla, “Beyond answers: Transferring reasoning capabilities to smaller llms using multi-teacher knowledge distillation,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.04616>
- [8] Y. Gu, L. Dong, F. Wei, and M. Huang, “Minillm: Knowledge distillation of large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.08543>
- [9] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, “A survey on model compression for large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2308.07633>
- [10] A. L. Tonja, B. F. Dossou, J. Ojo, J. Rajab, F. Thior, E. P. Wairagala, A. Aremu, P. Moiloa, J. Abbott, V. Marivate *et al.*, “Inkubalm: A small language model for low-resource african languages,” *arXiv preprint arXiv:2408.17024*, 2024.
- [11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.
- [13] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou, “A survey on knowledge distillation of large language models,” *arXiv preprint arXiv:2402.13116*, 2024.
- [14] P. Zhang, G. Zeng, T. Wang, and W. Lu, “Tinyllama: An open-source small language model,” *arXiv preprint arXiv:2401.02385*, 2024.
- [15] Y. Xu, X. Han, Z. Yang, S. Wang, Q. Zhu, Z. Liu, W. Liu, and W. Che, “Onebit: Towards extremely low-bit large language models,” *arXiv preprint arXiv:2402.11295*, 2024.
- [16] S. H. Muhammad, I. Abdulmumin, A. A. Ayele, N. Ousidhoum, D. I. Adelani, S. M. Yimam, I. S. Ahmad, M. Beloucif, S. M. Mohammad, S. Ruder *et al.*, “Afrisenti: A twitter sentiment analysis benchmark for african languages,” *arXiv preprint arXiv:2302.08956*, 2023.
- [17] M. R. Costa-Jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, “No language left behind: Scaling human-centered machine translation,” *arXiv preprint arXiv:2207.04672*, 2022.
- [18] W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Kolawole, T. Fagbohunbe, S. O. Akinola, S. H. Muhammad, S. Kabongo, S. Osei *et al.*, “Participatory research for low-resourced machine translation: A case study in african languages,” *arXiv preprint arXiv:2010.02353*, 2020.
- [19] K. Ogueji, Y. Zhu, and J. Lin, “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*, 2021, pp. 116–126.
- [20] O. Ogundepo, T. R. Gwadabe, C. E. Rivera, J. H. Clark, S. Ruder, D. I. Adelani, B. F. Dossou, A. A. Diop, C. Sikasote, G. Hacheme *et al.*, “Afriqa: Cross-lingual open-retrieval question answering for african languages,” *arXiv preprint arXiv:2305.06897*, 2023.
- [21] K. Ahmed and J. Buys, “Neural machine translation between low-resource languages with synthetic pivoting,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 12 144–12 158.
- [22] D. I. Adelani, J. Ojo, I. A. Azime, J. Y. Zhuang, J. O. Alabi, X. He, M. Ochieng, S. Hooker, A. Bukula, E.-S. A. Lee *et al.*, “Irokobench: A new benchmark for african languages in the age of large language models,” *arXiv preprint arXiv:2406.03368*, 2024.
- [23] J. S. Raju, J. S. Walia, S. Raghav, V. Marivate *et al.*, “Afroxlmr-comet: Multilingual knowledge distillation with attention matching for low-resource languages,” *arXiv preprint arXiv:2502.18020*, 2025.
- [24] Z. Liu, C. Zhao, F. Iandola, C. Lai, Y. Tian, I. Fedorov, Y. Xiong, E. Chang, Y. Shi, R. Krishnamoorthi, L. Lai, and V. Chandra, “Mobilellm: Optimizing sub-billion parameter language models for on-device use cases,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.14905>
- [25] J. O. Alabi, I. A. Azime, M. Zhang, C. España-Bonet, R. Bawden, D. Zhu, D. I. Adelani, C. O. Odoje, I. Akinade, I. Maab, D. David, S. H. Muhammad, N. Putini, D. O. Ademuyiwa, A. Caines, and D. Klakow, “Afridoc-mt: Document-level mt corpus for african languages,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.06374>