Swahili - LLM

Dr Michael Mollel (PhD)

Link to the project: GitHub
Link to the model: HF

Politics Part of LLM

DPO, KTO, RLHF etc all are complex things

Things to talk

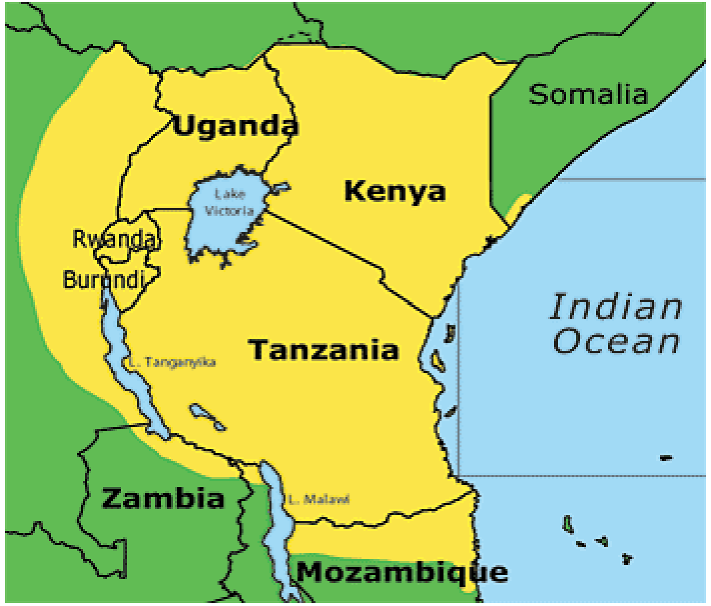
1. **Building the Bridge**
2. **Swahili LLM**
3. **What and Why**
4. **Empowering local content**
5. **End**

Bridging Language Gaps in AI

❖ Large Language Models (LLMs) like ChatGPT have revolutionized natural language processing (NLP) capabilities

❖ However, these models often struggle with low-resource languages and lack customization options

❖ Swahili, a widely spoken language in East and Southern Africa, has limited language support in current LLMs



- Swahili is spoken by more than 150 million people throughout the yellow-highlighted region
- Source: https://www.safaricrewtanzania.com/en/language-kiswahili/

Swahili-LLM

- Swahili –LLMs are the Large Language Model (LLM) that addresses the Swahili-speaking population's unique linguistic characteristics and needs.

The main objectives and features of Swahili-LLM include:

1. **Linguistic Inclusivity**: Bridging the gap in AI technology by providing a powerful tool for language processing and generation in Swahili, a language spoken by millions but often underrepresented in the digital and AI landscapes.
2. **Open-Source Model**: Unlike closed-source Large Language Models, which are restricted in access and customization, Swahili-Llama would ideally be open-source. This allows for greater flexibility, enabling developers, researchers, and users to tailor the model to their needs and contribute to its ongoing improvement.
3. **Versatile Application Support**: The model supports various applications, from text generation and sentiment analysis to supervised fine-tuning to different applications.
4. **Customization and Localization**: To offer an open-source solution that allows for extensive customization and localization, enabling developers, businesses, and governmental organizations to tailor AI applications to suit specific needs and cultural contexts of Swahili-speaking populations.

Swahili-LLM: Why Swahili LLM

- **Customization at Scale:** Explicitly tailored for Swahili, it allows for fine-tuning to meet the unique needs of various applications, from text generation to sentiment analysis.
- **Open-Source Accessibility:** Enables developers in low-resource settings to innovate and develop AI solutions that are genuinely relevant to their communities.
- **Bridging Language Gaps:** By providing a robust tool for Swahili, Swahili-LLM paves the way for greater inclusivity in AI, ensuring that technology serves everyone, irrespective of their primary language.

❖**The Impact:** Swahili-LLM is not just a technological breakthrough; it's a step towards democratizing AI, making it accessible and useful for millions of Swahili speakers. It represents a significant leap towards closing the digital divide, fostering local content creation, enhancing educational tools, and improving access to information.

Empowering Local Content and Businesses

- Swahili-LLM stands at the forefront of a transformative journey, breaking new ground by localizing AI technology for millions of Swahili speakers. Its development is pivotal in making advanced AI tools accessible and relevant to the East and Southern African communities.

❑**Key Use Cases:**

- **Small Business Support**: Swahili-LLM enables small businesses to leverage AI to enhance customer service, conduct market analysis, and create local content that resonates with their target audience.
- **Education**: By providing educational content in Swahili, the model can help overcome language barriers in learning, making education more inclusive and accessible.
- **Public Services**: Government and public service applications can use Swahili-LLM to develop sophisticated applications that ensure security and efficiency since it is open source.

Empowering Local Content and Businesses

❑**Driving Innovation and Inclusivity:**

- **Technological Advancement**: Swahili-LLM's sophisticated understanding of Swahili demonstrates a significant leap in AI, particularly in handling the language's nuances more efficiently than ever.
- **Community Empowerment**: This model empowers communities by enhancing digital literacy, fostering local content creation, and enabling a more inclusive digital economy.

❖**The Vision:** Our vision with Swahili-LLM is to create a tool and ignite a movement towards more linguistically inclusive technologies. We aim to inspire innovation that respects and uplifts local cultures and languages, making technology a true enabler of progress for all.

Technical Part of LLM

Prompting and RAG are equivalent to feature engineering (data augmentation, to be exact) in the traditional ML paradigm while finetuning and pretraining are equivalent to model training. So, all three are needed for the best results.

Andriy Burkov

TalentNeuron

Procedure for creating your LLMs

1. Vocabulary Expansion
2. Continue the Pretraining Foundation Model
3. Finetuning for Instruction Following
4. Finetuning for Model Alignment
5. Low-Rank Adaptation (LoRA)

Advance Finetuning

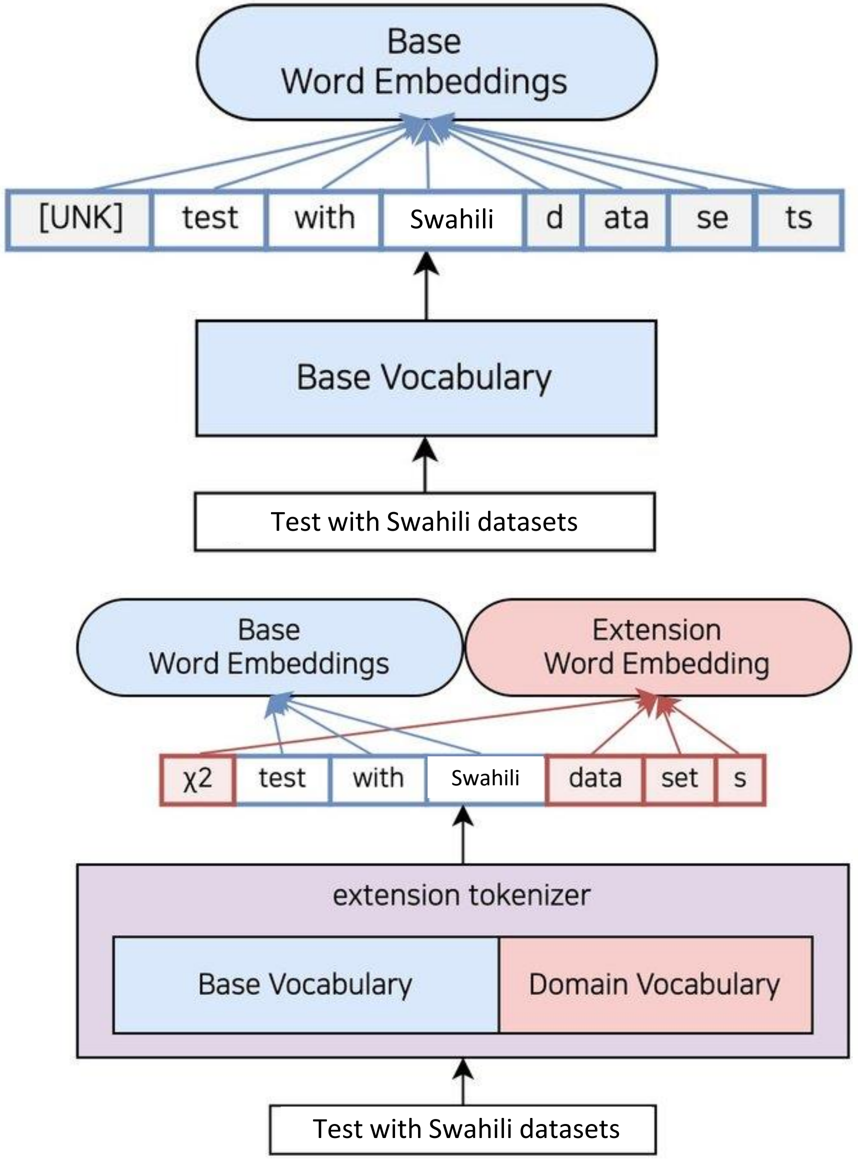
Vocabulary Expansion

a) Get your text corpus (Dataset approx. more than 1B words)

b) Train dataset with tokenizer model (**LLaMA 2, LLaMA3, Gemma – Sentence Piece Model**) link.

c) Merge the trained tokenizer with the original Foundation model tokenizer by taking the union of vocabularies

d) Resize LLM's word embeddings and language model head for the new merged vocabulary size



Base Word Embeddings

[UNK] | test | with | Swahili | d | ata | se | ts

Base Vocabulary

Test with Swahili datasets

Base Word Embeddings

Extension Word Embedding

χ2 | test | with | Swahili | data | set | s

extension tokenizer

Base Vocabulary

Domain Vocabulary

Test with Swahili datasets

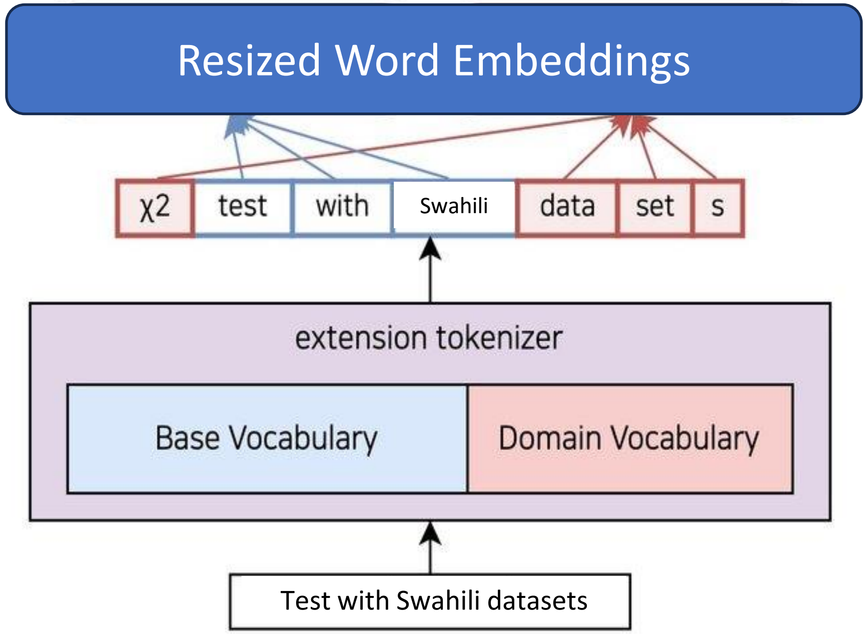
Vocabulary Expansion

a) Get your text corpus (Dataset approx. more than 1B words)

b) Train dataset with tokenizer model (**LLaMA 2, LLaMA3, Gemma – Sentence Piece Model**) link.

c) Merge the trained tokenizer with the original Foundation model tokenizer by taking the union of vocabularies

d) Resize LLM's word embeddings and language model head for the new merged vocabulary size

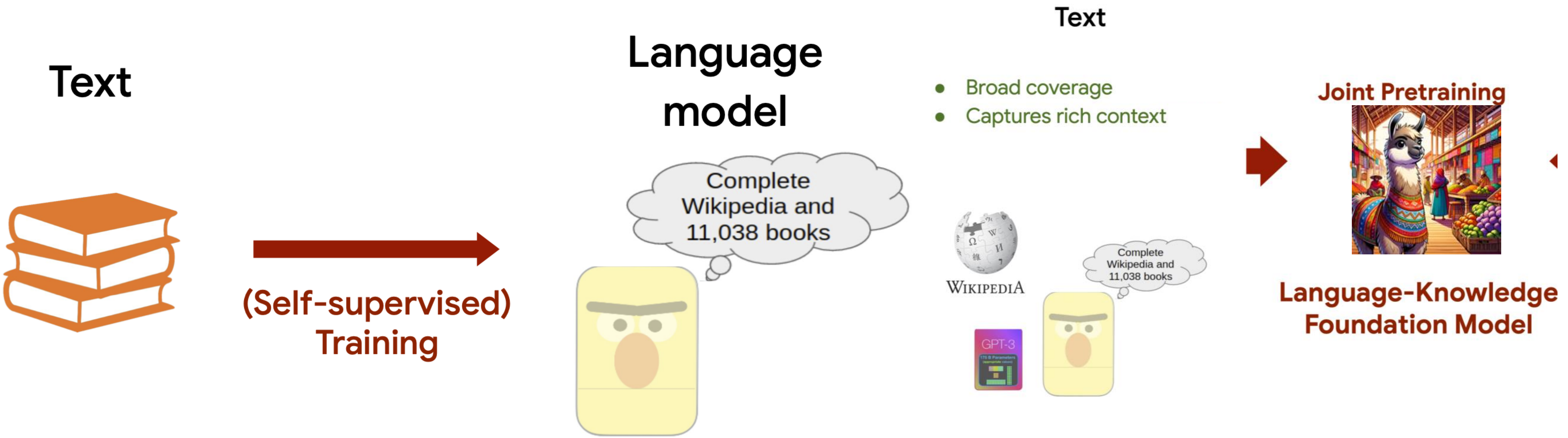


Continue the Pretraining Foundation Model

a) Initialize model with original Foundation model weights

b) Pretrain on Dataset corpus using Causal Language Modeling (CLM) as an objective

c) Apply Low-Rank Adaptation (LoRA) for efficient



Text

(Self-supervised) Training

Language model

Text

- Broad coverage
- Captures rich context

Joint Pretraining

Language-Knowledge Foundation Model

Complete Wikipedia and 11,038 books
WIKIPEDIA
GPT-3
Complete Wikipedia and 11,038 books

Finetuning for Instruction Following

a) Collect instruction-following data

b) Use prompt template:

```
### Maelezo:
{Majibu}
### Majibu:
{Majibu}
```

c) Finetune LLM on instruction data with supervised learning

d) Calculate loss only on the {Majibu} part of the sequence

e) Apply LoRA for efficient finetuning

Technique Finetuning

a) Collect instruction-following data

b) Use prompt template:

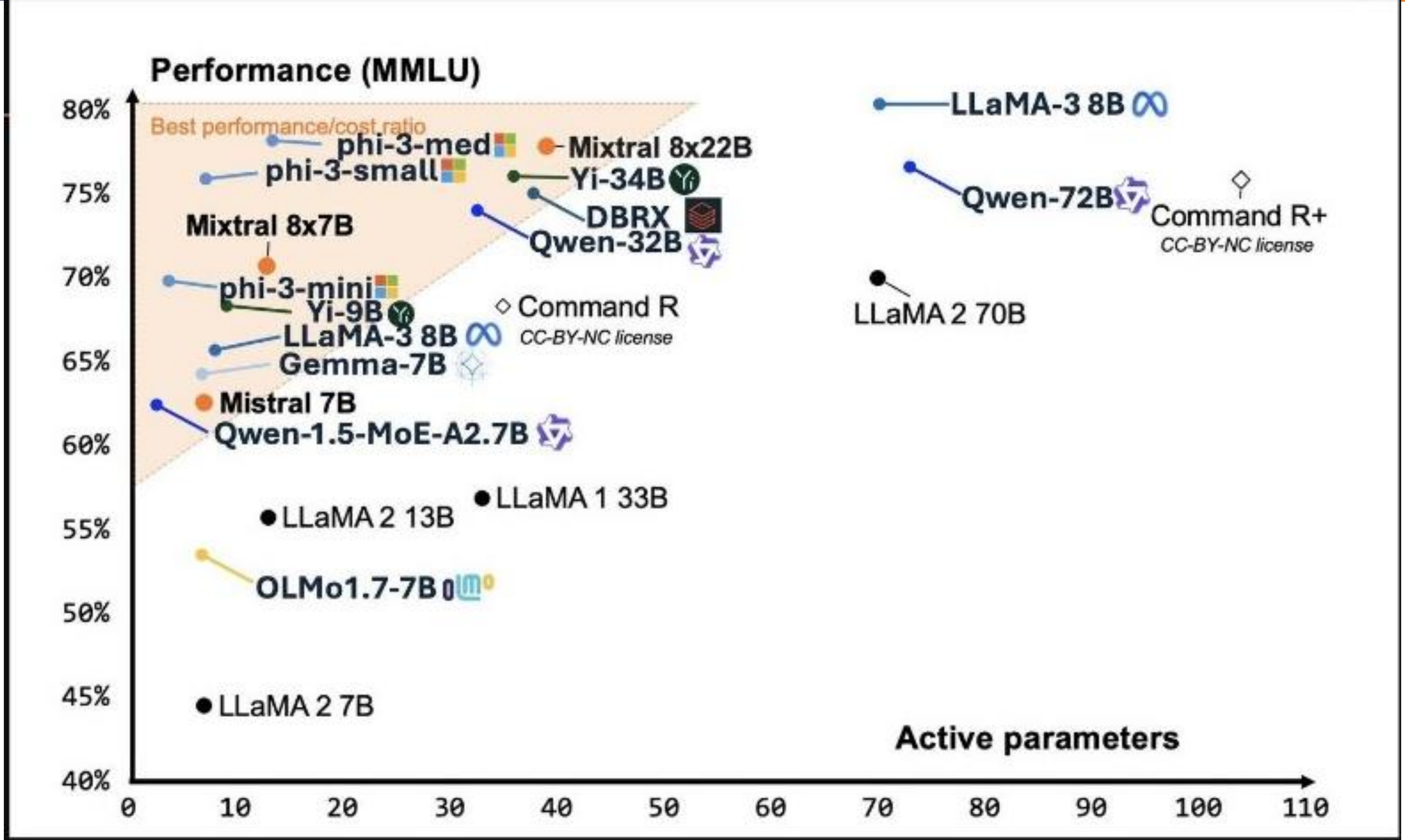
```
### Maelezo:
{Majibu}
### Majibu:
{Majibu}
```

c) Finetune LLM on instruction data with supervised learning

d) Calculate loss only on the {Majibu} part of the sequence

e) Apply LoRA for efficient finetuning

What are the best Foundation Model



Current Swahili Benchmark

Naïve Implementation (Unofficial)

S/N	Model Name	Average	ARC-C (300)	ARC-E	MMLU	HellaSwag	winogrande
1	Swahili –LLaMA-2 (7B)	<33%	<33%				
2	LLaMA-2 (7B)	<33%	<33%				
3	Kiswallama (7B)	<33%	<33%				
4	Ulizallama (7B)	<33%	<33%				
5	Swahili-Gemma (7B)	39%	39%				
6	Gemma (7B)	35%	35%				
7	Swahili-Alpaca-LLaMA3						
8	Swahili-LLaMA3						
9	LLaMA3						
10	Swahili-Alpaca-Phi3						
11	Swahili-Phi3						
12	Phi3						

Hackathon

Examples

- ❖Rag starter:

  https://www.kaggle.com/code/mikemollel/rag-gemma-swahili2.

- ❖SfT data creation example:

  https://www.kaggle.com/code/mikemollel/swahili-gemma-dataset-creation3.

- ❖Evaluate Swahili LLM ARC Challange Task:

  https://www.kaggle.com/code/mikemollel/evaluator-swahili-llms