Enhancing AI-Driven Farming Advisory in Kenya with Efficient RAG Agents via Quantized Fine-Tuned Language Models

Theophilus Owiti[1], **Andrew Kipkebut**[1]
[1]Kabarak University, Eldama Ravine Road, Kenya
Correspondence: lowiti@kabarak.ac.ke

Abstract

The integration of Artificial Intelligence (AI) in agriculture has significantly impacted decision-making processes for farmers, particularly in regions such as Kenya, where access to accurate and timely advisory services is crucial. This paper explores the deployment of Retrieval Augmented Generation (RAG) agents powered by fine-tuned quantized language models to enhance AI-driven agricultural advisory services. By optimizing model efficiency through quantization and fine-tuning, our aim is to deliver a specialized language model in agriculture and to ensure real-time, cost-effective and contextually relevant recommendations for smallholder farmers. Our approach takes advantage of localized agricultural datasets and natural language processing techniques to improve the accessibility and accuracy of advisory responses in local Kenyan languages. We show that the proposed model has the potential to improve information delivery and automation of complex and monotonous tasks, making it a viable solution to sustainable agricultural intelligence in Kenya and beyond.

1 Introduction

Despite Open-source and proprietary Pre-trained Language Models (PLMs) being trained on large sets of text data, they may lack fundamental principles, in each domain like Agriculture, which govern use of words to predict subsequent words in a sentence. This limitation is demonstrated by the fact that, most LLMs (Large Language Models) today have an exceptionally good global performance but fail in specific task-oriented problems (Josep, 2024). This research focuses on leveraging fine-tuning and quantization techniques for both small and large language models to create specialized models and agents for the agricultural sector in Kenya. While PLMs have demonstrated significant performance on downstream tasks for both high- and low-resourced languages, there is still a large drop in performance for underrepresented African languages during pre-training (Alabi et al., 2022).

Existing efforts, such as AfroLM, majorly focused on a multilingual language model pre-trained on 23 African languages using a self-active learning framework. Even though it performs well on various NLP downstream tasks like Named Entity Recognition (NER), text classification, and sentiment analysis (Dossou et al., 2022), these models lack sufficient linguistic and contextual grounding in Kenyan languages and agricultural knowledge.

In the plant health domain, studies have shown that PLMs are useful for text-mining applications but face challenges in low-resource settings due to limited labelled data. For example, research on text mining for plant health hazard detection highlights the need for models

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 24–30

trained on recent, domain-specific agricultural datasets (Jiang et al., 2023). To address these gaps, this research proposes fine-tuning existing PLMs on a corpus of Kenyan agricultural data, enhancing their ability to understand both domain-specific terminology and Kenyan languages. The goal is to develop a compact and efficient quantized language model optimized for agriculture, along with an AI agent capable of both comprehending Kenyan languages and executing task-oriented actions based on natural language inputs.

The absence of specialized agriculture models in Kenyan languages presents a significant gap. Kenyan farmers lack access to adequate and efficient AI-powered solutions that can provide up-to date, localized and contextually relevant agricultural information. Currently, available language models do not incorporate extensive agricultural expertise, nor are they optimized for Kenyan languages, making them ineffective for tasks such as farming guidance, risk assessment, financial literacy, and market insights.

This limitation demonstrates potentially dire consequences in the agricultural sector and Kenya's economy at large. Kenyan farmers struggle to access reliable and actionable farming information, from best planting practices to market trends and financial advice. The available resources are often generic, presented in English or Swahili, and fail to offer localized insights tailored to farmers' specific regions and crops. Sometimes farmers do not have adequate access to extension services. As a result, misinformation and a lack of accessible knowledge contribute to poor farming decisions, lower yields, and financial instability.

While AI-driven agricultural solutions in Kenya primarily focus on weather, soil analysis, crop disease detection, and pest control, they typically follow a three-step approach: detect a problem, offer recommendations, and direct farmers to Agro-vets for solutions. However, these solutions lack a unified access point, personalization and context-aware support that empowers farmers with continuous assistance. LLMs and AI agents have demonstrated their effectiveness in delivering instant, tailored information and executing actions in other domains, yet their potential remains untapped in Kenyan agriculture.

The purpose of the study is to develop a compact and efficient language model that understands Kenyan languages and agricultural terminology while integrating with AI agents to assist farmers. In doing so, our goal is to bridge the knowledge gap in agriculture, improve decision-making, and empower farmers with accessible and language-inclusive AI support.

2 Related Work

Adapting Pre-trained Language Models for African Languages Several efforts to use pretrained models have led to multilingual fine-tuning approaches for African languages. One of the most effective approaches to adapt to a new language is language adaptive fine-tuning (LAFT) — fine-tuning a multilingual PLM on monolingual texts of a language using the pretraining objective. However, adapting to a target language individually takes a large disk space and limits the cross-lingual transfer abilities of the resulting models because they have been specialized for a single language. They performed multilingual adaptive fine-tuning in 17 most resourced African languages and three other high-resource languages widely spoken on the African continent to encourage cross-lingual transfer learning (Alabi et al., 2022). AfroLM a multilingual language model pretrained from scratch on 23 African languages (the largest effort to date) using our novel self-active learning framework. Pretrained on a dataset significantly (14x) smaller than exist-

ing baselines, it outperforms many multilingual pre-trained language models (AfriBERTa, XLMR-base, mBERT) on various NLP downstream tasks like NER and text classification (Dossou et al., 2022).

Tool-calling Other models outperform Text-Davinci-003 and Claude-2, achieve comparable performance to ChatGPT, and is only slightly inferior to GPT4. Besides, models (ToolLLaMA) exhibits robust generalization to previously unseen APIs, requiring only the API documentation to adapt to new APIs effectively. They majorly focus on Supervised fine-tuning to enhance tool calling capabilities with synthesized training data (Qin et al., 2023).

3 Standard and Instruction-based Chain of Thought Annotation

Current LLMs also exceed in areas such as tool calling and reasoning with chain-of-thought (CoT). CoT instruction tuning has drawn attention for its potential to encourage complex, step-by-step reasoning. LLMs can demonstrate CoT abilities with proper prompting and instruction engineering (Liu et al., 2023), but this is something that lacks in most parts of Africa given the state of underrepresented African languages.

Tool-calling is a focus area in this research where we aim to achieve a model that can support tool calling in local Kenyan languages. To achieve this, we focus on adopting Supervised fine-tuning (SFT). It is a method to enhance the tool calling capabilities of LLMs, with the training data often being synthesized. The current data synthesis process generally involves sampling a set of tools, formulating a requirement based on these tools, and generating the call statements (Wang et al., 2024).

Data based on local Agriculture documents have been created with paired translations from vernacular languages to English. We employ an instruction-based format in the dataset, heavily focusing on tuning the training data to employ a Chain of Thought format and focus on tool calling annotation for single-tool calling in Kenyan languages, with standard prompts we focus on enhancing chat completion capabilities based on the corpus that has been translated to specific native Kenyan languages.


```
{
  "api_list": [
    {
      "category_name": "maize",
      "tool_name": "plant_diagnosis",
      "api_name": "CheckPlantHealth",
      "api_description": "Checks for crop health in maize.",
      "required_parameters": ["location", "crop_type"],
      "method": "GET"
    }
  ],
  "query": "Apidho ga oduma, to adak Kisumu, Kenya. Odumba gi nitie gi kumoro ma rateng'. Ango
ma chamo gi?",
  "relevant APIs": [
    [
      "CheckPlantHealth"
    ]
  ],
  "query_id": 1
}
```


Figure 1: Example of the tool calling dataset used with queries in Kenyan languages. This particular training data uses Dholuo.

Construction of the tool-calling dataset for training is split into three stages: collecting existing APIs and creating missing APIs spanning across different categories (such as Agribusiness, Weather, News for Farmers), writing instructions in given languages covering APIs for single-tool scenarios, and the solution path annotation for each instruction (Qin et al., 2023). An example of this is found in Figure 1.

4 Method

To develop a multilingual model specialized in the agricultural domain, we adopted the following steps to achieve the goal of creating a model specialized for a subset of African languages and measuring how well a fine-tuned and quantized PLM can perform in various agricultural tasks in Kenya.

4.1 Data Collection

Data was collected from disparate sources. For instance, the language pair sentences provided by Tech Innovators Network Kenya (THiNK), publicly available websites and documents about agriculture and biblical text extracted using OCR. The THiNK dataset comprised of local-language and swahili pairs[1], therefore, to have a local-language and english pair of each language (Luyha, Luo, Kalenjin, Kidaw'ida-Kiswahili) the dataset comprising of 91,097 sentence pairs was changed to accomodate the manual translation of Swahili words to English. See Table 1 for an example of the Language pair for local-language and swahili sentence pairs.

Language Pair	Train Set Size	Test Set Size	Total Size (bytes)
Kidaw'ida-Kiswahili (dav_swa)	21,329	5,333	1,973,706
Kalenjin-Kiswahili (kln_swa)	28,101	7,026	3,537,847
Dholuo-Kiswahili (luo_swa)	23,446	5,862	4,387,588

Table 1: Total number of language pair sentences provided by THiNK.

4.2 QLoRA for Finetuning

For efficient fine-tuning, we propose using QLoRA (Quantized model weights + Low-Rank Adapters) with the models Llama-2-7B, Llama-3-8B Instruct and Llama-3.1-8B, which reduces memory usage without compromising a model's initial performance as demonstrated in Figure 2. This fine-tuning technique offers several advantages: 4-bit NormalFloat, which outperforms 4-bit Integers and 4-bit Floats in empirical results; Double Quantization, which compresses quantization constants, saving an average of 0.37 bits per parameter; and Paged Optimizers, which mitigate memory spikes caused by gradient checkpointing when processing long-sequence mini-batches (Dettmers et al., 2023).

This approach is informed by the adoption of Quantile Quantization, an optimal data type that estimates the quantile of the input tensor using the empirical cumulative distribution function (Dettmers et al., 2023). While quantization of LLMs has traditionally focused on inference, QLoRA has demonstrated a breakthrough by enabling backpropagation through frozen, quantized weights at large model scales (Belkada et al., 2023).

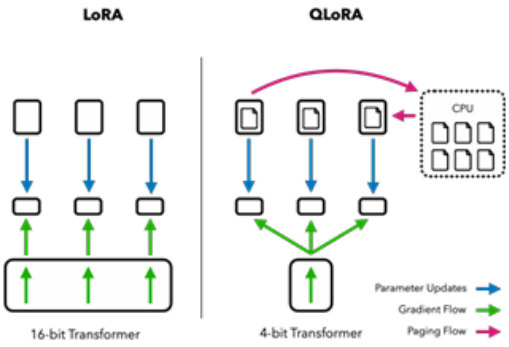

Figure 2: QLORA proves to be efficient than LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes. the image is taken from (Dettmers et al., 2023)

4.3 Supervised Fine-tuning

SFT (Supervised Fine-tuning), adapts a pre-trained model to a specific task by taking labelled datasets as input constructed for intended tasks. To be effective, a significant amount of raw data and resources are required to construct and label SFT datasets (Ross et al., 2025).

The research aims to use refined Africa corpora from various languages in Kenya. These data will be further added to standard and chain-of-thought instruction sets with translation pairs from English to local Kenyan languages. During this process of SFT the

[1] https://huggingface.co/datasets/thinkKenya/kenyan-low-resource-language-data

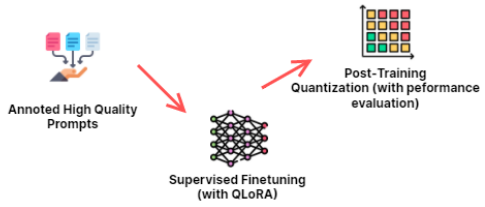

Figure 3: A brief workflow of the training process. This summarizes how the annotated language data is passed over for training to QLoRA and further subjected to testing.

data will then be fed to the training stage where we will use Transformer Reinforcement Learning (TLR)[2] that provides SFTTrainer that makes it straightforward to supervise fine-tune open LLMs, it is a subclass of the Trainer from the transformers library and supports all the same features, including logging, evaluation, and checkpointing, but adds additional quality of life features, including PEFT (parameter-efficient fine-tuning) support including Q-LoRA, or Spectrum (Schmid, 2025) This stage has been indicated on Figure 3.

4.4 Post-Training Quantization

This involves taking our fine-tuned model and quantizing the model parameters during the inference phase. This method does not involve any changes to the training process itself. The dynamic range of parameters is recalculated at runtime, like how we worked with the example matrices (Valenzuela, 2024). This technique allows reducing the size of these increasingly the fine-tuned models with an aim of making it perform better at Agriculture but can also be easily allowed to run on consumer-grade devices with minimal performance depreciation. The quantized model will then be subjected to further testing based on various agricultural areas and the performance will be evaluated and tuned further to achieve optimal results.

[2] https://huggingface.co/docs/trl/index

4.5 RAG Agents and Tool-calling

In this research, we utilize LangGraph to develop an external agent and create Agentic RAG applications that enhance the decision-making process of the deployed model. These applications enable the model to determine whether to retrieve information from the vector store or generate responses directly, as illustrated in Figure 4. Additionally, we extend the agent's capabilities by leveraging LangChain and LangGraph to orchestrate a fine-tuned open-source model deployed on Hugging Face. This allows farmers to interact with the system using natural language to retrieve relevant information and seamlessly access various automated tools for diverse functions.

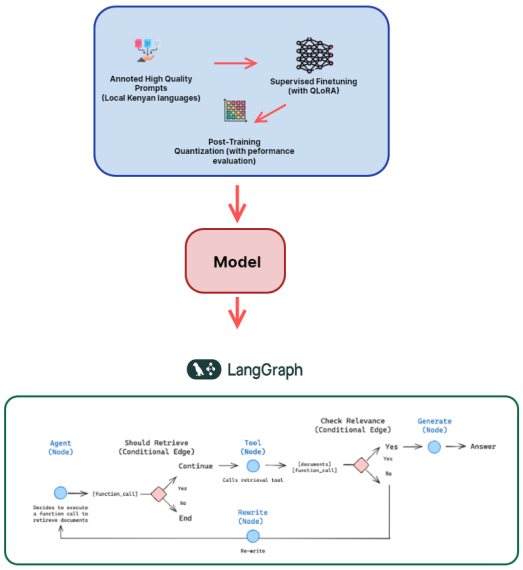

Figure 4: Implementation of the fine-tuned model for AI Agents using LangChain

Beyond this, we will evaluate the in-built too-calling to determine its performance and effectiveness. The capability of tool-calling

is achieved by the annotated data set that had CoT format instructions that improves reasoning and also the trained model benefits from the instruction-format based data given during training.

5 Conclusion

This study demonstrates the potential of quantized fine-tuned language models in improving AI-driven farming advisory services through efficient RAG agents. By optimizing model size and computational efficiency, we enable real-time, localized, and cost-effective recommendations tailored to the needs of smallholder farmers in Kenya. Our findings indicate that the proposed approach enhances response accuracy and system performance compared to conventional models, reducing resource constraints while maintaining high-quality advisory outputs. Future work will focus on expanding dataset coverage, integrating multimodal inputs such as satellite imagery, and refining model interpretability to further enhance AI-driven agricultural decision-making in resource-limited environments.

Limitations

Fine-tuning methods utilized in this study have proven to be effective, although, a significant limitation is the scarcity of high-quality, annotated datasets available for fine-tuning models in African languages. The process of creating such datasets is resource-intensive and time-consuming, requiring extensive hours of data preparation. This scarcity of readily available data has created challenges in developing robust and accurate models tailored for the agricultural sector in Kenya and other African contexts.Africa faces a significant shortage of diverse language corpus datasets that authentically capture the nuances of communication in its indigenous languages. To address this, we propose the development of comprehensive, high-quality datasets that reflect the linguistic diversity and cultural contexts of African indigenous languages, enabling more accurate and inclusive natural language processing applications.

Acknowledgments

This research would not have been possible without the exceptional support of our stakeholders. Special thanks to our mentors and colleagues for their valuable insights, feedback, and discussions, which greatly contributed to refining our work. We also recognize the immense support from various open-source communities and developers whose tools, datasets, and frameworks played a crucial role in this study. Furthermore, we appreciate the farmers and agricultural experts in Kenya who shared their experiences, helping us better understand the practical needs of AI-driven farming advisory.

References

O. J. Alabi, I. D. Adelani, M. Mosbach, and D. Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 2022 Conference on Computational Linguistics (COLING)*.

Y. Belkada, M. Sun, T. von Köller, S. Mangrulkar, B. Bossan, L. Debut, and S. Liu. 2023. Finetune llms on your own consumer hardware using tools from pytorch and hugging face ecosystem.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

F. P. Dossou, L. A. Tonja, Y. Oreen, S. Osei, A. Oppong, S. Iyanuoluwa, and C. 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. Association for Computational Linguistics.

S. Jiang, S. Cormier, A. Rafael, and F. Rousseaux. 2023. Improving text mining in plant health domain with gan and/or pre-trained language model. *Frontiers in Artificial Intelligence*.

F. Josep. 2024. Fine-tuning llms: A guide with examples.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023. Logicot: Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2908–2921.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

E. Ross, Y. Kansal, J. Renzella, A. Vassar, and A. Taylor. 2025. Supervised fine-tuning llms to behave as pedagogical agents in programming education. *arXiv preprint arXiv:2502.20527*.

Philipp Schmid. 2025. Fine-tune llms in 2025. Accessed: 2025-03-06.

Andrea Valenzuela. 2024. Quantization for large language models (llms): Reduce ai model sizes efficiently. Accessed: 2025-03-06.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Toolflow: Boosting llm tool-calling through natural and coherent dialogue synthesis. *arXiv preprint arXiv:2410.18447*.