Sauti AI - Technical Roadmap

Parameter-Efficient Fine-tuning of Gemma 3 1B for African Languages

From Swahili to Kikuyu: A Proven, Reproducible Framework for African Languages

Submission Date: 30th January 2026
Competition Timeline: 30th January - 20th March 2026

Executive Summary

This technical roadmap documents a practical, production-ready approach for developing culturally authentic conversational AI for African languages through parameter-efficient fine-tuning (PEFT). I have developed, deployed, and validated a dual-interface Swahili conversational AI system that achieves a validation perplexity of 3.97, demonstrating a substantial improvement over existing Swahili models, which typically report validation perplexities in the range of 20 to 45. The two-stage methodology combines continued pretraining on a diverse twelve-million-token corpus with targeted conversational fine-tuning using LoRA and 4-bit quantization, completing the full training process in approximately nine hours on accessible GPU infrastructure. During the competition period, I will demonstrate this methodology's reproducibility by repeating the complete training pipeline while establishing the foundation for systematic expansion to Kikuyu. This approach transforms one time achievements into a verifiable, scalable framework for other Kenyan (and African) languages. My work directly addresses critical gaps in narrative depth, dialogue patterns, and cultural nuances that have limited previous African language AI systems. It also mitigates issues with cross-lingual transfer, where models reason abstract concepts in Western languages and produce imperfect or unnatural replies in swahili and other African languages. I am driven by the vision of “AI that thinks the African way”, a call first articulated at the Africa Premier AI Conference(APAIC) in August 2025.

1. Current Status

1.1 Operational Swahili MVP Completed

I have developed and deployed a production quality Swahili conversational AI that achieves exceptional performance through my two-stage methodology. The final model demonstrates strong training stability with a training loss of 1.080 and validation loss of 1.387, showing excellent generalization. My system architecture employs CraneAILabs/swahili-gemma-1b with parameter-efficient LoRA fine-tuning, delivering production-ready performance on accessible hardware.

The Two Stage Method: My continued pretraining phase enhanced the model's foundational Swahili competence, enhanced cultural context understanding, better handling of complex grammatical structures, and stronger baseline performance before conversational fine-tuning. The subsequent conversational fine-tuning achieved exceptional dialogue-specific capabilities with smooth convergence within just seven epochs.

Qualitative Excellence: Through native speaker collaboration and careful corpus curation, I aimed to capture qualitative enhancements including genuine cultural authenticity from traditional folktales and Zanzibar narratives, linguistic sophistication from translated literary works, and improved dialogue quality with better coherence in multi-turn conversations. My expert validated translation methodology contributed sophisticated dialogue patterns, subtext, metaphorical expressions and cultural nuances while maintaining cultural authenticity.

Reproducibility: During the competition period, I will publicly re-execute my entire training pipeline from scratch, documenting every step to prove that my approach consistently produces these SOTA results. This commitment to reproducibility underscores the engineering rigor of my framework, distinguishing it from one time experiments.

1.2 Critical Lessons from Implementation

My development process revealed several critical insights that will guide future expansions. My two-stage methodology proved essential, as extensive pretraining created a robust foundation that dramatically improved conversational fine-tuning convergence. My literary translation approach, using GPT-4 for initial translation followed by expert refinement, produced higher quality training data than typical web scraped content. Swahili expert involvement prevented culturally inappropriate patterns, while my parameter-efficient approach, combining LoRA with 4-bit quantization, enabled single-GPU training without sacrificing quality.

Translation Methodology: I developed and validated a rigorous translation pipeline for dialogue-rich literary works, producing the first-ever Swahili translations of several major public-domain classics. The process begins with an initial GPT-4 translation from English, followed by expert refinement in close collaboration with native Swahili speakers and linguists. Subsequent cultural adaptation ensures tonal fidelity, contextual appropriateness, and idiomatic naturalness, with quality assurance enforced through multiple review cycles. This

methodology yielded historically novel Swahili translations of Anton Chekhov's *The Duel* (psychological dialogue), Leo Tolstoy's *The Death of Ivan Ilyich* (philosophical and existential discourse), and Robert Louis Stevenson's *The Strange Case of Dr. Jekyll and Mr. Hyde* (narrative and dramatic dialogue), each contributing uniquely to conversational depth and stylistic range within the training corpus.

Technical Configuration details: Through systematic testing, I established optimal configurations including LoRA with rank 8 and alpha 16, targeting the q_proj, v_proj, k_proj, and o_proj modules with a dropout rate of 0.05, while utilizing only approximately 0.5% of the base model parameters. My use of 4-bit NF4 quantization with double quantization enabled successful training on accessible hardware while maintaining model quality.

1.3 Framework Ready for Replication

I have developed and validated a complete, replicable framework for systematic language expansion. The two-stage training methodology has been thoroughly tested through the Swahili implementation. The entire data pipeline, from literature acquisition through translation, validation, and preprocessing, has been fully operationalized and optimized. The infrastructure has also been proven sufficient, with a single P100 GPU successfully handling the full training process in approximately nine hours.

My evaluation framework combining quantitative metrics with native speaker and expert validation provides robust assessment capabilities. My deployment architecture featuring a dual-interface system (Kiswahili RAG Learning Assistant and General Conversation Assistant) has been implemented and tested, demonstrating practical utility. My quality assurance process involving multi-stage native speaker review ensures consistent cultural and linguistic appropriateness.

This framework is ready for immediate application to Gikuyu with kikuyu data. Based on my Swahili experience, I estimate 6-8 weeks from corpus completion to production deployment for Kikuyu, with each subsequent language benefiting from my accumulated refinements.

2. Proven Architecture & Methodology

2.1 Two-Stage Training Methodology (Validated & Reproducible)

My systematic two-stage methodology has proven highly effective for building comprehensive African language capabilities in LLMs:

2.1.1 Continued Pretraining

The first stage focuses on continued pretraining to establish a strong foundation. I employed a 12 million toke corpus drawn from diverse sources, training for two epochs over six hours on a Kaggle P100 GPU. This stage fosters deep internalization of cultural and linguistic patterns often missed by other approaches.

2.1.2 Conversational Finetuning

The second stage specializes in conversational fine-tuning, cultivating natural and culturally appropriate dialogue capabilities. For Swahili, this process achieved a validation perplexity of 3.97, with training completed in approximately 2.5 hours. Convergence was smooth within seven epochs, reflecting stable learning dynamics built upon the robust pretraining foundation.

This two-stage separation ensures that the model develops strong linguistic understanding during pretraining while fine-tuning refines conversational skills without catastrophic forgetting. Overfitting on the limited dialogue dataset is minimized, and cultural grounding occurs primarily in Stage 1, with dialogue specialization handled in Stage 2. I will demonstrate the reproducibility of this methodology during the competition through a fully documented and publicly accessible replication of both stages.

2.2 Base Model Specification

I selected Crane AI's swahili-gemma-1b, a Swahili extended variant of Gemma-1B, featuring approximately one billion parameters and a tokenizer adapted for Swahili-specific tokens. This model provides a linguistically rich starting point while remaining computationally accessible, allowing deployment on consumer-level hardware with fast inference for real-time conversation. The choice balances capability with efficiency, and my production deployment has validated that high-quality conversational AI can be achieved with modest parameter counts when combined with targeted, high-quality training data.

2.3 Parameter-Efficient Training Strategy

My parameter-efficient training employs LoRA (Low-Rank Adaptation) with rank 8, alpha 16, targeting Query, Value, Key, and Output projection layers with dropout 0.05. This configuration uses trainable parameters representing only approximately 0.5% of the base model's total parameters. I selected this configuration through systematic testing as it provided optimal performance for this application, enabling single-GPU training on free-tier hardware while maintaining quality.

I implemented 4-bit NF4 quantization with double quantization, achieving 4x memory reduction with minimal quality loss. This enables loading the full model

on a 16GB GPU while using FP16 mixed precision for gradient computation. This strategy has been validated through my Swahili implementation, showing no measurable quality degradation while dramatically reducing resource requirements.

2.4 Computational Efficiency & Accessibility

My two-stage resource utilization demonstrates exceptional computational efficiency: continued pretraining required 6 hours total on Kaggle GPU P100 using LoRA, while conversational fine-tuning required approximately 2.5 hours on a single GPU P100 with LoRA. Total development time was approximately 9 hours across both stages, with training speed of 0.29 iterations/second during fine-tuning. Memory optimization through 4-bit quantization combined with LoRA efficiency enabled successful training on accessible hardware.

2.4.1 Infrastructure & Software Stack

The computational backbone relies primarily on Kaggle's P100 GPU with 16GB VRAM, available through their free tier. The project requires approximately 50GB of storage for datasets, models, and checkpoints, along with stable internet connectivity for dataset downloads and model uploads.

The software stack builds on the Transformers library from Hugging Face, combined with PEFT for parameter-efficient fine-tuning and Bitsandbytes for quantization. PyTorch 2.0 or later provides the deep learning framework, while Weights & Biases or TensorBoard handles experiment tracking and visualization.

This computational accessibility democratizes access to high-quality language model development for low-resource languages. My successful use of 4-bit quantization and LoRA demonstrates that advanced NLP capabilities can be developed without extensive computational infrastructure, making this methodology accessible for African language AI development initiatives.

3. Data Strategy & Corpus Development

3.1 Continued Pretraining Corpus I Built (Swahili)

I have already curated a comprehensive 12 million token Swahili corpus through a multi-source approach emphasizing quality and cultural relevance. The foundation includes public domain Swahili literature: The Swahili Bible, *Kasiri ya Mwinyi Fuad* by Adam Shafi Adam for classical Swahili prose blending moral instruction with cultural tradition, *The Complete Works* of Shaaban Robert which are landmarks of modern Swahili literature, *Utengano* by Said Ahmed Mohamed exploring societal struggles, *Walenisi by* Katama Mkangi as speculative Swahili

literature with political allegory, and *Zanzibar Tales* for traditional folklore and oral traditions.

The inclusion of classic, public-domain literary works was undertaken to address a specific deficiency in existing low resource language corpora. What makes this project unique is its deliberate focus on building a training corpus that captures narrative arcs, sustained dialogue patterns, metaphor, and subtext. These elements are critical for training high quality language models, yet they are largely absent from most Swahili datasets. Narrative structure teaches a model how meaning unfolds over time by guiding emotional and logical progression, while subtext encodes nuance, implication, and unspoken intent. In the absence of these features, language models tend to generate outputs that are shallow, repetitive, and stylistically limited.

My translated literary classics employed rigorous methodology for dialogue-rich works. I included *The Duel* by Anton Chekhov for psychological dialogue, *The Death of Ivan Ilyich* by Leo Tolstoy for philosophical discourse, *Dr. Jekyll and Mr. Hyde* by R.L. Stevenson for narrative dialogue, and *The Communist Manifesto* for political discourse.

Additional sources included Swahili Wikipedia for encyclopedic knowledge, government documents for formal register, and news archives from BBC Swahili for contemporary usage. This corpus achieves high domain diversity across literature, politics, culture, news, and folklore, with language registers spanning formal, informal, literary, and conversational styles.

The continued pretraining corpus was standardised using a structured text format designed to preserve document (one complete, self-contained piece of text that you feed to the model as a single unit) boundaries, topical context, and semantic continuity. Each document is encoded with an explicit title marker followed by the main body text, using a consistent schema of **<|title|>** to denote the document heading and **<|text|>** to introduce the content, with the model's native **<eos>** token marking the end of a document and sequence. The **<bos>** token precedes the entire sequence to signal the beginning of the document.

For example, an educational passage is represented as **<bos><|title|>Fasihi Simulizi<|text|>** followed by the full prose content and ending with **<eos>**.

3.2 Dataset Creation via Omnilingual ASR

Beyond traditional text collection, I am integrating **Meta's Omnilingual ASR** (released November 2025) to enable authentic spoken Kikuyu content transcription. This approach addresses a critical challenge in low-resource language AI: the scarcity of written digital content despite rich oral traditions.

Corpus Development Applications:

- **Oral Tradition Preservation**: Transcribing Kikuyu folktales, proverbs, and community narratives from audio recordings
- **Conversational Data**: Capturing authentic dialogue patterns from natural Kikuyu speech
- **Cultural Content**: Converting spoken performances, storytelling sessions, and interviews into training data

This methodology enriches my training corpus with conversational cadences and idiomatic expressions that rarely appear in written text, providing linguistic authenticity that purely text-based approaches cannot achieve.

3.3 Conversational Dataset for Instruction Tuning (3,175 Examples)

Available conversation datasets, such as the Swahili Stanford Alpaca dataset, contain significant noise, and poor-quality data produces poor models(GIGO). To address this, I curated a modest but high-quality conversational dataset of 3,175 examples, with 2,857 allocated for training and 318 for validation. The dataset includes both single-turn and multi-turn dialogues in a user-assistant format, manually curated for linguistic accuracy and cultural appropriateness. Preprocessing follows Gemma-specific formatting with <start_of_turn> and <end_of_turn> tokens, enforces a maximum sequence length of 512 tokens, and ensures a consistent dialogue structure. This dataset provides a solid foundation and can be scaled up in future expansions.

3.4 Kikuyu Corpus Development

My Kikuyu corpus development targets 4 to 6 million tokens for Stage 1 pretraining and 1,500 conversational examples for Stage 2 fine-tuning, applying my proven methodology. I am collecting Kikuyu oral traditions, contemporary literature, and dialogue-rich classics for translation, supported by native speaker validation. The translation pipeline follows my established process, with GPT-4 providing initial translations, followed by native speaker refinement and cultural adaptation across multiple review cycles. Additionally, I have discovered a novel approach to dataset creation for low-resource languages using Meta's omnilingual ASR (released November 2025), enabling accurate transcription of spoken Kikuyu content from youtube channels which enriching the corpus with authentic conversational data.

4. Results & Performance Validation

4.1 Quantitative Performance

My two-stage methodology achieved a final validation perplexity of 3.97, representing a 32.3% reduction from epoch 1. The model converged effectively within seven epochs, demonstrating training stability with minimal overfitting (Δloss = 0.307). The training progression showed consistent improvement: epoch 1 (5.91), epoch 2 (5.07), epoch 3 (4.52), epoch 4 (4.19), epoch 5 (4.04), epoch 6 (4.08), reaching 3.97 at epoch 7.

This performance establishes new standards for Swahili language modeling, where existing models typically range between 20-45. My results demonstrate effective capture of Swahili linguistic patterns and conversational structures through my comprehensive approach.

4.2 Dual Chat Assistant System

To demonstrate practical applications, I developed a dual-chat interface system featuring a unified backend with a single model serving two specialized interfaces, a modular design that allows independent prompting strategies and knowledge integration, and scalable web-based deployment capable of supporting concurrent users:

4.2.1 Kiswahili Learning Assistant (RAG Pipeline):

The system serves educational purposes, supporting high school Kiswahili language learning (Forms 1 to 4) through a RAG pipeline built on a curated Swahili knowledge base aligned with the national curriculum. It features confidence scoring to ensure reliability, and as research shows that students understand concepts more effectively when taught in their mother tongue, the ultimate goal is to develop an educational assistant capable of teaching a range of subjects entirely in Kiswahili.

RAG Implementation

1. **Vector database:** Kiswahili high school curriculum.
2. **Retrieval mechanism:** Semantic similarity search.
3. **Confidence scoring:** Retrieval relevance assessment.

4.2.2 General Conversation Assistant:

The system supports open-domain conversation using direct model generation guided by prompt engineering, relying on the model's intrinsic knowledge. Its primary use cases include creative tasks, casual chat, and general-purpose

dialogue, enabling engaging and versatile interactions beyond structured educational content.

4.2.3 Speech-to-Text Integration

To maximize accessibility for speakers of various Kenyan languages and content creators, I will integrate Meta's Omnilingual ASR into my dual-chat interface, enabling speech-to-text in the following languages: Swahili, Kikuyu, Kamba, and Luhya, Luo, Kalenjin, and Maasai. Users speak in these languages and Sauti AI (that's the name of my project) transcribes it to text. YouTube content creators in these languages will also be able to upload their audio/video and the system generates accurate subtitles.

4.3 Cultural & Linguistic Contributions

My systematic inclusion of traditional Swahili literature, folktales, contemporary works, and the first Swahili translations of world literature created a comprehensive linguistic foundation that respects both historical and modern Swahili expression. This approach addresses critical gaps in narrative depth, dialogue patterns, and cultural nuances that have limited previous Swahili AI systems. By integrating continued pretraining with carefully curated literature, expert-validated translations, and diverse cultural texts, I established a robust foundation for specialized conversational capabilities. Collaboration with native speakers and linguists ensured cultural authenticity while introducing sophisticated dialogue patterns, enabling the model to capture the richness of Swahili language and culture.

5. Competition Timeline & Deliverables

ACTIVITY / MILESTONE	DEADLINES						
	Week 1 (Jan 30–Feb 5)	**Week 2 (Feb 6–12)**	**Week 3 (Feb 13–19)**	**Week 4 (Feb 20–26)**	**Week 5 (Feb 27–Mar 5)**	**Week 6 (Mar 6–12)**	**Week 7 (Mar 13–19)**
Swahili Corpus Review & Sharing							
Re-execute Swahili Stage							

1 Pretraining & share logs							
Swahili Conversational Fine-Tuning & share logs (Stage 2)							
Dual-Chat Interface Development & ASR integration							
Kikuyu Corpus Collection & Preprocessing							
Kikuyu Initial Pretraining (Stage 1)							
System Testing & QA for Swahili Assistant							
Final MVP Submission							

Milestone 1: Technical Roadmap (30th January 2026)

This comprehensive technical roadmap represents my first deliverable, formalizing my methodology and establishing my reproducibility commitment.

Milestone 2: Data & Development Environment (13th February 2026)

I will share my custom Swahili corpus and also initiate and document a complete re-execution of my Swahili training pipeline under observable conditions using my exact configurations, while concurrently advancing Kikuyu corpus development to 40-50% completion.

Milestone 3: Functional Alpha (27th February 2026)

I will present my fully operational Swahili system with both interfaces, complete my reproducibility demonstration showing final metrics matching my results, and demonstrate methodology transfer to Kikuyu through initial pretraining progress.

Milestone 4: System Security & Integration (13th March 2026)

I will implement production-grade security and optimization for my Swahili system while designing and documenting a multi-language architecture capable of supporting simultaneous deployment of multiple language models.

Milestone 5: Complete MVP (20th March 2026)

I will present a complete MVP package including professional video demonstration, comprehensive project report with all validation data, impact metrics dashboard, user testimonial compilation, live system access, reproducibility proof documentation, and future roadmap.

6. Limitations & My Future Directions

I acknowledge several current limitations that inform my future development:

Dataset and Scope Constraints:

My conversational fine-tuning relied on 3,175 examples, which provides a great starting point, though larger dialogue datasets could further enhance performance. Also, the pretraining corpus, while diverse, reflects selected literary and cultural sources, and geographic or dialectical variations across Swahili-speaking regions are not comprehensively represented. Additionally, domain-specific applications may require further specialized fine-tuning to achieve optimal results.

Translation and Cultural Considerations:

Although my translated literary works underwent expert validation, they may still contain subtle cultural transfer artifacts. Striking the right balance between preserving the original literary intent and adapting for cultural authenticity requires ongoing refinement. Additionally, contemporary Swahili digital communication patterns are only partially represented, leaving room for future expansion to capture modern usage more fully.

Technical and Evaluation Constraints:

While the evaluation relied primarily on quantitative metrics`, comprehensive human assessment across diverse use cases is still needed to fully capture conversational quality and cultural appropriateness. Long-term performance in real-world deployments remains to be observed, and applying the methodology to other African languages will be important to verify its generalizability, ensuring that both linguistic accuracy and cultural authenticity hold across different language contexts.

Future Research Directions:

1. Expansion to other Kenyan languages, beginning with Kikuyu, using the proven two-stage methodology.

2. Larger pretraining corpora integrating additional Swahili literary works, historical documents, and contemporary digital content.

3. Conversational dataset expansion beyond 3,175 examples

4. Domain specialization through development of specialized corpora for medical, financial and agricultural applications.

5. Expanded literary translation projects with continued expert validation.

6. Multimodal integration, including speech recognition .