BULaMU: An Open Foundation Model for Luganda

Mwebaza Rick
ricky.mwebaza@gmail.com

**Introduction**

Uganda, colloquially referred to as the “pearl of Africa”, is home to over 50 million people, 56 tribes, and over 40 languages, most of which descend from the Bantu family (Namyalo et al., 2016).

Luganda is the most popular language in Uganda, boasting over 20 million speakers around the globe (Indiana University). Around 70% of all the speakers of Luganda reside within the nation's borders, mostly occupying the Baganda area (Indiana University).

This paper introduces BULaMU (**B**reakthrough in **U**tilization of **L**arge L**a**nguage **M**odels in **U**ganda), the first Large Language Model that has been trained from scratch in Luganda. This model is both free and open weights and will be available on Huggingface to download for others to experiment with or adapt to their use case.

**Related Work**

Advanced computing runs deep in Africa, with algorithms like the Ifa Divination system, which is based on complex binary code that conceptually underpins many modern machine learning classifiers (Alamu et al., 2013).

Local talent, in collaboration with companies like Samasource, has played pivotal roles in the development of groundbreaking technologies like the Xbox Kinect Sensor (ZDNET, 2019), which would eventually lay the foundation for the development of Apple’s Face ID (Coldewey, 2017). A few months ago, Felix Kitaka provided a demonstration for “LugandaGPT” on his Youtube Channel, the first voice assistant in Luganda (Atino, 2025).

Uganda has been making noticeable strides in broadening access to computing within the nation, especially since access to computing remains a challenge for the majority of the population. 28 percent of the population has access to the internet (Miharia, 2025), with only 9% of people over the age of 10 using the internet over the last year (Watchdog Uganda, 2025). This figure is up 13% from 2019, indicating the number of people with access to the internet has nearly doubled in 6 years (World Bank, 2025). Uganda initiated Vision 2040 (Miharia, 2025) to continue broadening access to computing by pledging to train 3 million Ugandans on how to use AI and creating 1 million AI centered jobs by 2030 (United Nations Uganda, 2025). This plan has been aided by the deployment of the continent’s “first AI factory” in Uganda in partnership with NVIDIA (Okafor, 2025). Several groups are working on fine tuning other models.

**Methods**

BULaMU was trained on a non-machine translated, 390M token corpus that was created by leveraging data that is publicly available on the internet as well as datasets from sources like Wikipedia, Huggingface, GitHub, Common Crawl. The data was cleaned, deidentified, and deduplicated using an ensemble of python scripts.

BULaMU is a compute-optimal, autoregressive, decoder only transformer model with 6 layers, 6 attention heads, hidden size of 384, and a context length of 384. The base model of BULaMU was trained for 20000 iterations on Apple Silicon (M4 chip, 10 core CPU, 10 core GPU,  16 GB ram) . The base model had a train loss and validation loss of 2.6883. The fine-tuned model had a train and validation loss of 2.589. I fine-tuned the model for 3000 iterations on Apple Silicon on a 102,000 token dataset of publicly available question-answer pairs and natural language inference tasks that were translated to Luganda using Google Translate. I wanted to see whether this would improve its performance on Natural language inference (NLI) problems compared to the base model. I used the problems from IrokoBench to conduct my evaluation.

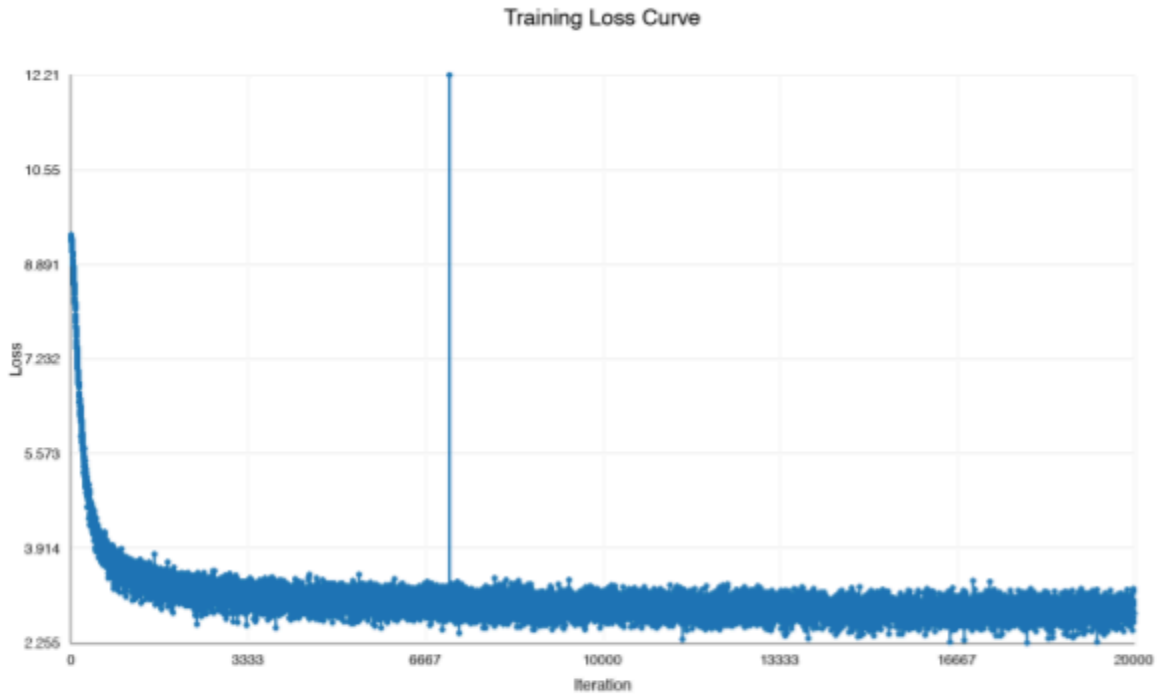


**Figure 1:** Shows the val/loss (training loss) curve for the BULaMU base model

**Results**

I evaluated the NLI capabilities of the base and fine tuned versions of BULaMU using the 600 "test" examples from IrokoBench. I wrote a python script to determine how many examples the both models were able to label and how accurate the labels were. The python script looks for the first digit that comes after "label:" in the model's output.

| | BULaMU-base | BULaMU-instruct |
|---|---|---|
| Number of Examples | 600 | 600 |
| Number Labeled | 303 | 475 |
| Number Unlabeled | 297 | 125 |
| Percentage Labeled | 50.50% | 79.17% |
| Number Correct | 67 | 136 |
| Percent Correct | 22.11% | 28.63% |

**Figure 2:** Compares the performance of the base and instruct model on NLI tasks

The base model only labeled 50.5% (297) of the examples while the finetune model was able to label 79.17% (125) of the examples from IrokoBench. Instruction fine tuning also improved the accuracy of BULaMU by 6.52% from the base model.

**Discussion**

This paper introduces BULaMU, the world's first foundation model that has been trained from scratch in Luganda. BULaMU illustrates that even "tiny", compute-optimal, resource efficient language models are able to understand the fundamental language structure of Luganda, follow instructions, and make their own inferences based on the data they were trained and finetuned on. BULaMU also demonstrates that fine tuning greatly improves the model's ability to complete specialized tasks in Luganda.

BULaMU's performance may significantly benefit from several improvements. BULaMU has not been evaluated for safety and may reflect some of the biases in its training dataset. The dataset, while comparatively large for a "low-resource language", is still significantly smaller than the corpora used to train language models for high resource languages, like Common Crawl, OpenWebText, or BookCorpus. Machine Translation (MT), particularly from Google Translate or Google Gemini, may offer a partial solution to this issue but should be approached with extreme caution. The quality of MT is still poor and may not reflect the speech patterns that native speakers exhibit (Mwebaza, 2025). This may have limited the performance of the instruct version of BULaMU, which was finetuned using publicly available datasets that were translated to Luganda.

BULaMU's performance may have also been limited by its low parameter count, as language models with more parameters have been noted to have better "grammatical judgement" (Dentella et al., 2025). More research needs to be done on how scaling laws (Kaplan, Hoffmann, etc.) apply to Luganda, since there is conflicting evidence on how these principles apply to low-resource languages. Researchers in one paper found that in low-resource scenarios, the scaling laws relating to model size, data, and compute do not necessarily apply (Urbizu et al., 2023). In fact, they found that a larger, undertrained model outperforms a smaller, compute optimal-model (Urbizu et al., 2023).

I will be making BULaMU available to the public for download under the Apache 2.0 license because I believe that it can enable community-driven AI solutions in Uganda. Its small size significantly reduces the barrier to entry for using AI, as it should be possible to inference the model on low-power, low-cost devices like cell-phones, laptops, or embedded systems. I am going to continue to develop strong, tiny language models, which have around 100M parameters or fewer, that run on edge computing devices (Guertler et al., 2024). Moving forward, I may also develop a larger model to leverage the similarity that Luganda has to other Ugandan and Bantu Languages. Many Ugandans are multilingual (Westbrook, 2022) and are able to code-switch between multiple languages during conversation (Otundo and Muhleisen, 2022).

## References


Aashna Miharia. (2025, July 16). *How an International NGO Plans to Fight the Digital Divide with a Satellite Company - Non Profit News | Nonprofit Quarterly*. Non Profit News | Nonprofit Quarterly. https://nonprofitquarterly.org/how-an-international-ngo-plans-to-fight-the-digital-divide-with-a-satellite-company/

Alamu F.O, & Ho, A. (2021). *A COMPARATIVE STUDY ON IFA DIVINATION AND COMPUTER SCIENCE* (W. I. Isharufe, Ed.). https://www.semanticscholar.org/paper/A-COMPARATIVE-STUDY-ON-IFA-DIVINATION-AND-COMPUTER-F.O-Ho/024464da04bb7739cbb6f7cd42c74a377e04fb85

Atino, V. (2025, January 19). *First AI-powered Luganda Voice Assistant Developed in Uganda.* Nilepost News. https://nilepost.co.ug/news/237940/first-ai-powered-luganda-voice-assistant-developed-in-uganda

Coldewey, D. (2017, September 12). *iPhone X basically has a Kinect on the front to enable Face ID | TechCrunch*. TechCrunch. https://techcrunch.com/2017/09/12/iphone-x-basically-has-a-kinect-on-the-front-to-enable-faceid/

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Agerri, R., & Aitor Soroa. (2023). *Scaling Laws for BERT in Low-Resource Settings*. https://doi.org/10.18653/v1/2023.findings-acl.492

Hillier, D., Guertler, L., Tan, C., Agrawal, P., Ruirui, C., & Cheng, B. (2024). Super Tiny Language Models. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.14159

Luganda. (2021). *Luganda*. Center for Language Technology. https://celt.indiana.edu/portal/Luganda/index.html

Mwebaza, R. (2025). *Lost in Translation: Evaluating the Performance of Five Publicly Available Machine-Translators for Luganda-English Interpretation*. SocArxiv. https://doi.org/10.31235/osf.io/j36fs_v1

Okafor, C. (2025, September 11). *Another "first AI factory in Africa"? Uganda is said to have made its entry into the AI race*. Business Insider Africa. https://africa.businessinsider.com/local/markets/another-first-ai-factory-in-africa-uganda-is-said-to-have-made-its-entry-into-the-ai/n47mrec

Otundo, B., & Mühleisen, S. (2022). Code-switching and advising in multilingual African situations: An analysis of radio phone-in programmes in Kenya and Cameroon. *Journal of the Language Association of Eastern Africa*, *1*(1), 1–15. https://doi.org/10.5642/jlaea.cbof4616

Vittoria Dentella, Günther, F., & Leivada, E. (2025). Language in vivo vs. in silico: Size matters but Larger Language Models still do not comprehend language on a par with humans due to impenetrable semantic reference. *PLoS ONE*, *20*(7), e0327794–e0327794. https://doi.org/10.1371/journal.pone.0327794

Watchdog Uganda. (2024, October 23). *REPORT: Only 13 percent Ugandan internet users use it for business and the rest for social chats*. Watchdog Uganda. https://www.watchdoguganda.com/news/20241023/173582/report-only-13-percent-ugandan-internet-users-use-it-for-business-and-the-rest-for-social-chats.html

Westbrook, J., Baleeta, M., Dyer, C., & Islei, A. (2022). Re-imagining a synchronous linguistic landscape of public and school uses of Runyoro-Rutooro and Runyankore-Rukiga in

early childhood education in Western Uganda. *Journal of Multilingual and Multicultural Development*, 1–14. https://doi.org/10.1080/01434632.2022.2038181

ZDNET. (2019, February 1). *Building AI data sets in Uganda | ZDNet*. YouTube. https://www.youtube.com/watch?v=MYahV4wZteA