Bridging Zambia's Wikipedia Knowledge Gap using Hybrid Retrieval Augmented Generation and Large Language Models

Frazer Nyambe [1,2] and Lighton Phiri [1,2]
{frazer.nyambe,lighton.phiri}@cs.unza.zm
[1]Department of Computing & Informatics, School of Natural & Applied Sciences, University of Zambia
[2]DataLab Research Group, University of Zambia

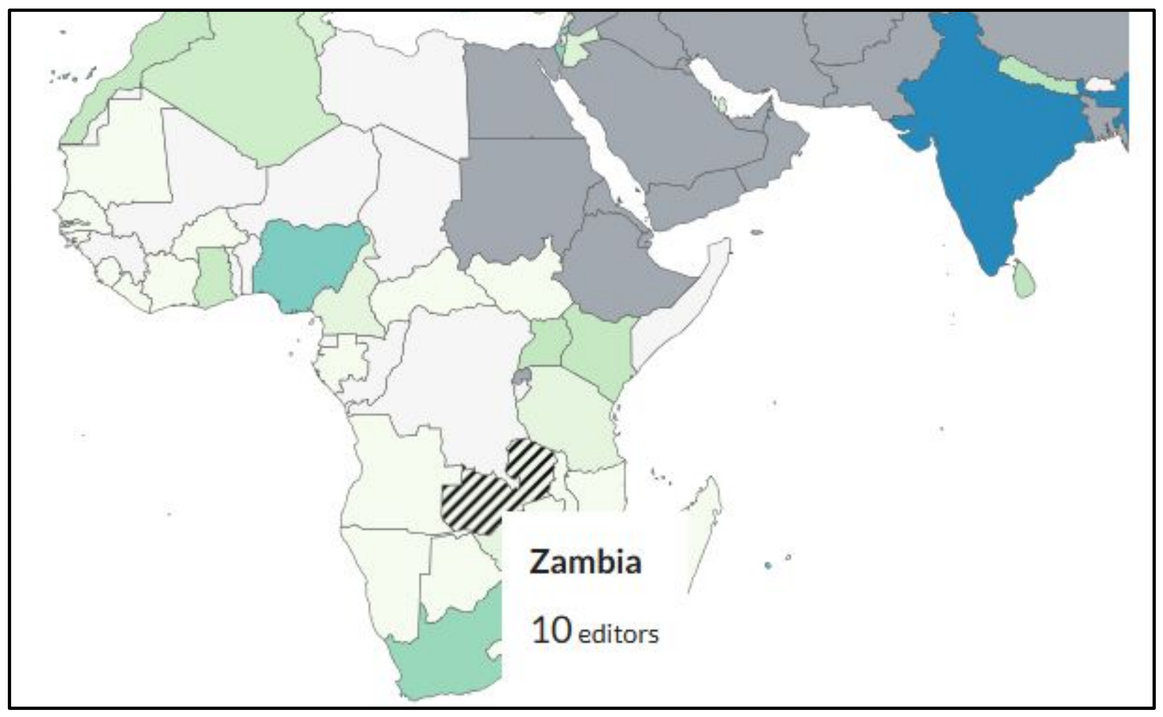
Introduction

Wikipedia has low local participation in Zambia despite high content consumption [1] resulting in gaps in Zambian knowledge. LLMs can support content creation, but hallucinations and poor factual grounding remain challenges [2]. This study explores base LLMs, fine-tuned LLMs, and Hybrid RAG approaches to improve Zambia-specific content generation.

Problem Statement

Low Zambian information and unreliable AI-generated information limit the development of accurate local Wikipedia content [3]. The study therefore compares three AI workflows to determine their ability to produce accurate, grounded and complete content and improve human editors' efficiency and experience.

Figure 1. Wikipedia Content Contribution in Zambia



Research Objectives

1. To design and configure three distinct local human–AI content generation workflows within an Open WebUI ecosystem consisting of a base open-source LLM, a fine-tuned domain-adapted LLM, and a Hybrid RAG pipeline using a curated Zambia-specific dataset.
2. To develop an evaluation framework for benchmarking the three AI configurations based on factual accuracy, text grounding, faithfulness, and contextual recall.
3. To experimentally evaluate the impact of the optimized AI-assisted platform on student editors' editing efficiency, content expansion depth, and user experience.

Methodology

This study adopts a quantitative approach using an experimental research design to develop and evaluate three software tools. The research is structured around the following specific objectives:

Objective 1. Design and configure three content generation workflows

- (i) OpenWebUI was integrated with the pre-trained Llama 3.1; (ii) Zambian data collected and used to fine-tune Llama 3.1 using QLoRA; (iii) Hybrid RAG prototype implemented in Open WebUI, using a vector-based knowledge base.

Objective 2. Evaluation Framework for the three content generation workflows

- The three systems will be evaluated based on factual acuracy, text grounding and contextual recall of produced output after a user prompt.

Objective 3. Experimentally Evaluation of the three content generation workflows

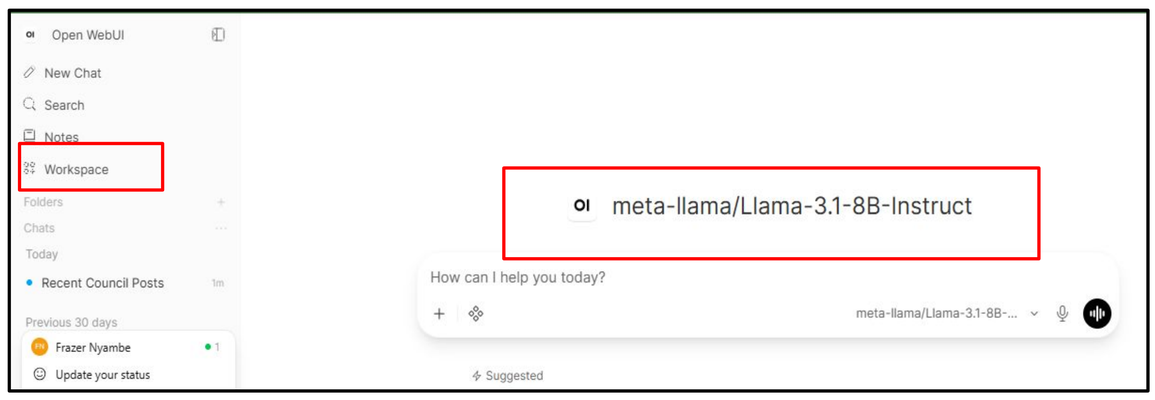
- A counter-balanced controlled within-subjects experiment will be conducted to evaluate the effect of the optimised platform. Participants will edit a Wikipedia article using the three tools Base LLM, Fine-Tuned LLM, and Hybrid RAG system and subsequently complete a questionnaire evaluating their experience with each tool.

Preliminary Findings and Discussion

Objective 1.1. Base open-source LLM

- The Base LLM uses the pre-trained Llama 3.1 model without Zambia-specific fine-tuning or external retrieval.
- The editor enters a topic/prompt, which is sent to Llama 3.1 to generate Wikipedia-style content based on its existing knowledge. The generated content is then reviewed, verified, and edited by the human editor.

Figure 2. Open WebUI Implemented with Pre-trained Llama3



Objective 1.2. Fine tuning open-source LLM
Fine-tuned dataset comprised of information from Zambian universities and Local councils, 13 million out of 8 billion parameters were modified using QLoRA

Figure 3. Fine Tuning Llama 3.1

```
# Start training
print("Starting fine-tuning...")
trainer.train()
print("Fine-tuning complete!")

# Save the fine-tuned model and tokenizer
save_path = "./fine_tuned_llama_model"
trainer.save_model(save_path)
tokenizer.save_pretrained(save_path)
print(f"Fine-tuned model and tokenizer saved to {save_path}")
```

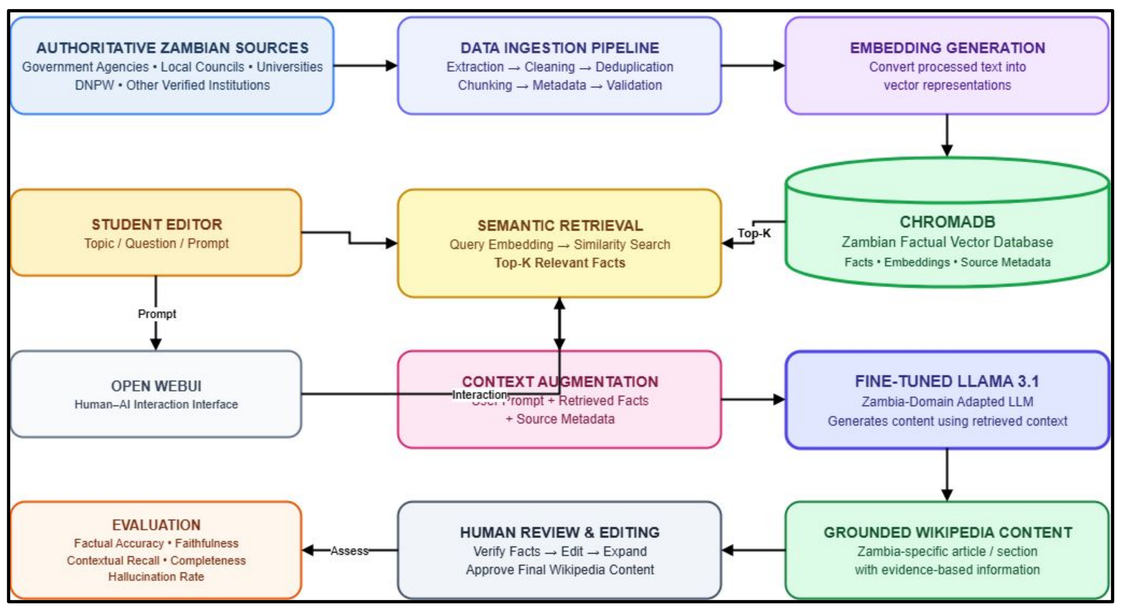
```
LoRA model created successfully.
trainable params: 13,631,488 || all params: 8,043,892,736 || trainable%: 0.1695
Starting fine-tuning...
```

Preliminary Findings and Discussion

Objective 1.3. A Hybrid RAG System Architecture

The Open WebUI RAG prototype demonstrates the feasibility of localized document ingestion and context-aware generation, with results showing that optimized chunking and vector search reduces hallucinations and improved response accuracy when using internal data.

Figure 4. Hybrid Retrieval Augmented Generation Architecture



Bibliography

[1] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S. and Li, Q. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. DOI: 10.48550/arXiv.2405.06211
[2] Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L. and Shazeer, N. 2018. Generating Wikipedia by Summarizing Long Sequences. DOI: 10.48550/arXiv.1801.10198
[3] Thompson, B., Dhaliwal, M., Frisch, P., Domhan, T. and Federico, M. 2024. A Shocking Amount of the Web is Machine Translated: Insights from Multi-way Parallelism. Findings of the Association for Computational Linguistics ACL 2024 (Stroudsburg, PA, USA, 2024), 1763–1775. DOI: 10.48550/arXiv.2401.05749

Contact Information

DataLab Research Group
University of Zambia
Private Bag 32379
Lusaka, Zambia 10101

Web: datalab.unza.zm
Email: datalab@unza.zm

DataLab Research Group