



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

School of Natural and Applied Sciences

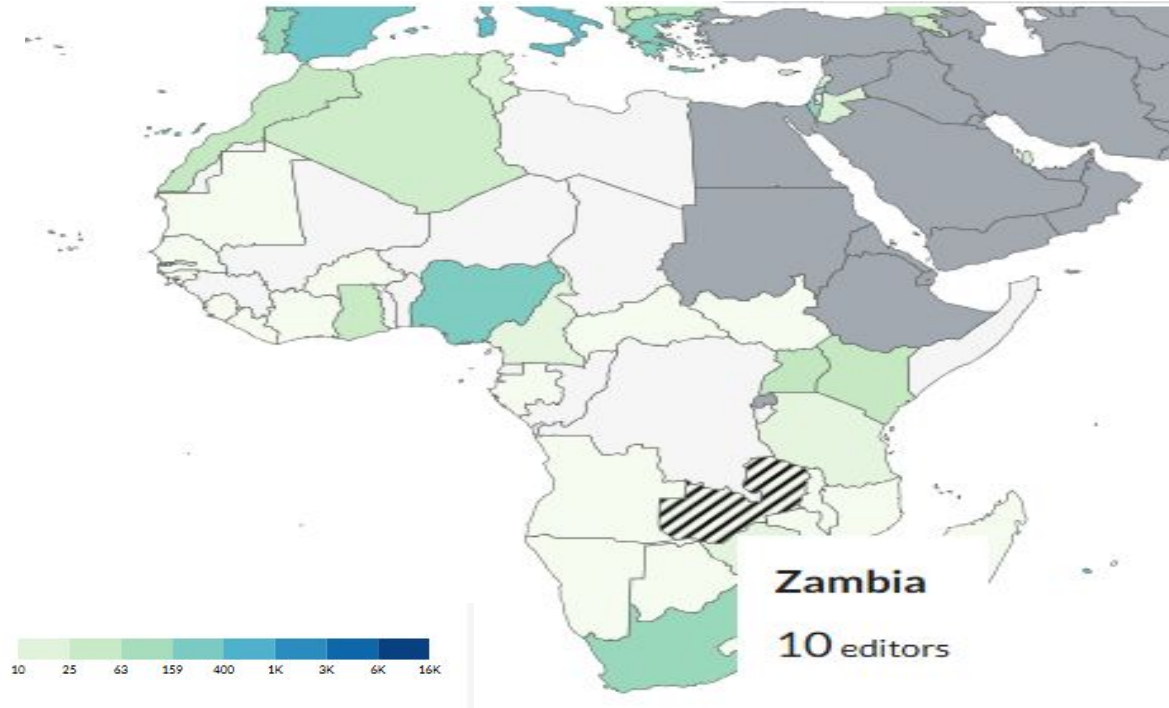
Department of Computing and Informatics

Presenter: Frazer Nyambe  
Computer: 2024020539

Supervisor: Dr. Lighton Phiri



# Wikipedia Content Contribution in Zambia



# Introduction

## Background

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

## Problem Statement

- ❑ Low Zambia information and unreliable AI-generated information limit the development of accurate local Wikipedia content.
- ❑ 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.

# Introduction

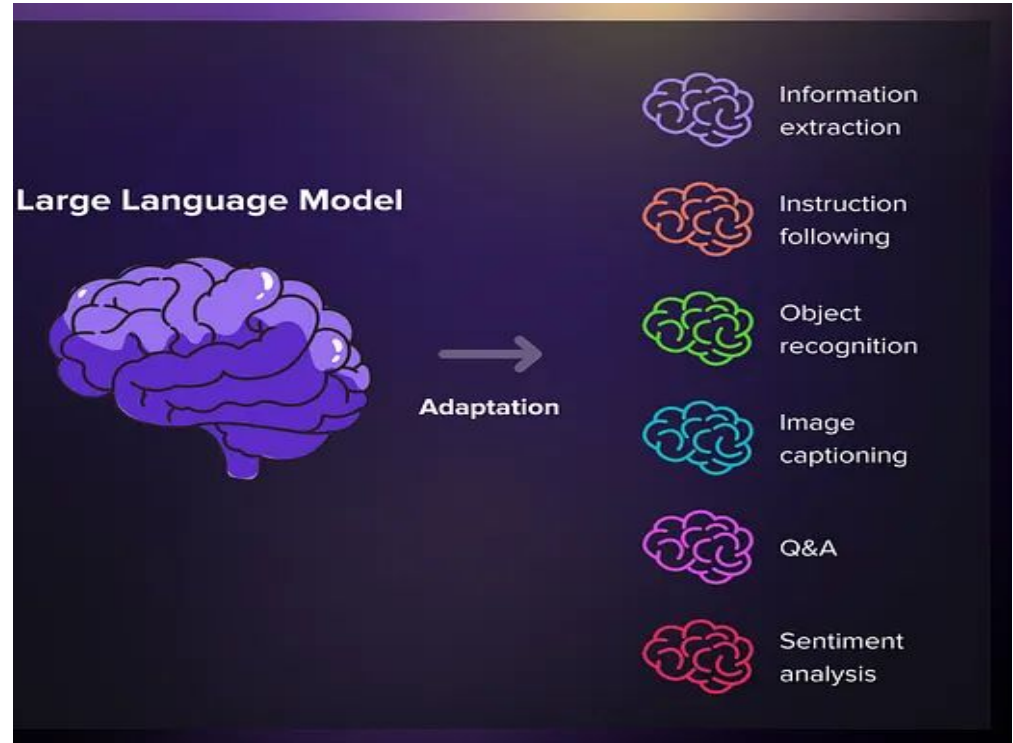


## Purpose of study:

- ❑ This study demonstrates how LLMs can be used to generate information about Zambia from grounded Zambia-specific factual sources.

## Significance of study:

- ❑ Enable Wikipedians to edit content about Zambia on wikipedia.



# Research Objectives and Research Questions



| Research Objective                                                                                                                                                                                                                                             | Research Question                                                                                                                               |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| 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. | 1. How can a local Open WebUI ecosystem integrate base LLM, fine-tuned LLM, and RAG workflows for Zambia-specific Wikipedia article generation? |
| 2. To develop an evaluation framework for benchmarking the three AI configurations based on factual accuracy, text grounding, faithfulness, and contextual recall.                                                                                             | 2. Which AI configuration provides superior performance in factual accuracy, grounding, 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.                                                                                           | 3. How does the optimized AI-assisted platform influence student editors' productivity and experience during Wikipedia article development?     |

# Introduction

## ❑ Operational Definitions

- ❑ **Generative AI:** AI systems capable of creating new content, such as text.
- ❑ **RAG (Retrieval-Augmented Generation):** AI approach combining document retrieval and language generation to ensure factual grounding.
- ❑ **Vector Database:** A semantic search system storing documents as vectors to retrieve contextually relevant data.

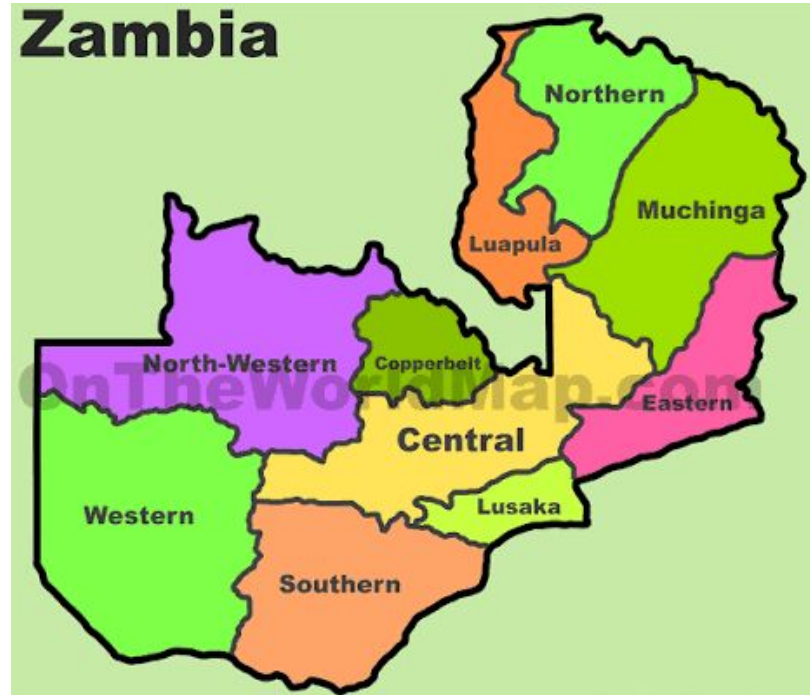
## ❑ Limitations of Study

- ❑ Controlled environment for the user test experiment
- ❑ Limited focus of the study due to limited publicly available data



# Zambia's Administration

- ❖ Zambia has over 113 councils managed by the Ministry of local government authorities, comprised of 5 city, 15 municipal, and 93 town councils.
- ❖ Most Zambian councils have Wikipedia pages and their information is on the official site for the Ministry of local government.



# The Higher Education Authority Statistics

The Higher Education Authority (HEA) is a statutory body that regulates and monitors standards in higher education institutions to ensure high-quality services and national human capital development



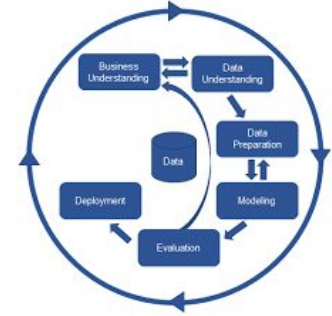
# Literature Review



1. Fan et al. (2024) showed that Retrieval-Augmented Generation (RAG) enhances generative AI by retrieving relevant external knowledge during generation.
2. Liu et al. (2024) found that combining extractive and abstractive summarization techniques can support Wikipedia article generation by improving the processing of lengthy input and output content.
3. Thompson et al. (2024) demonstrated that machine-translated and low-quality web content can negatively affect the reliability of generative AI systems.

# Methodology [Overview]

- ❑ **Research Design Approach:** Quantitative Research
- ❑ **Process Framework:** This study adopted the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) model to guide the design and implementation of the RAG-based system
- ❑ The study is aligned with the three specific objectives to enable *data triangulation* for improved, valid and reliable findings.



## 1 Design and Configuration of the three AI Workflows

Base open-source LLM, a fine-tuned domain-adapted LLM, and a Hybrid RAG pipeline using a curated Zambia-specific dataset.

## 2 Develop Evaluation framework for Benchmarking the three AI workflows

Factual accuracy, text grounding, faithfulness, and contextual recall

## 3 Human Evaluation of the three Human-AI workflows

The impact of the optimized AI-assisted platform on student editors' editing efficiency, content expansion depth, and user experience.

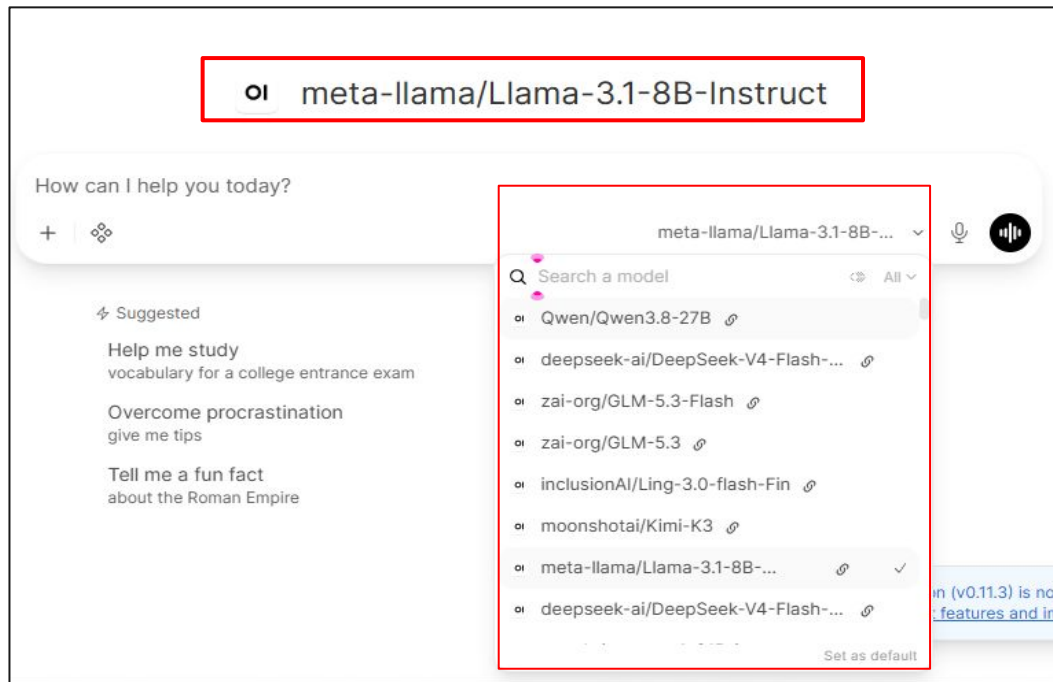
# Methodology [1/6]

## Objective 1: Design and configure three human-AI content-generation workflows

### A. Base open-source LLM

- Integration of OpenWebUI with the pre-trained Llama 3.1.

**Note:** Open WebUI supports various LLM runners like Ollama and OpenAI-compatible APIs, with built-in inference engine for LLMs



# Methodology [2/6]

## Objective 1: Design and configure three human-AI content-generation workflows

### Why use LLM like Llama-3 as a Base Open source?

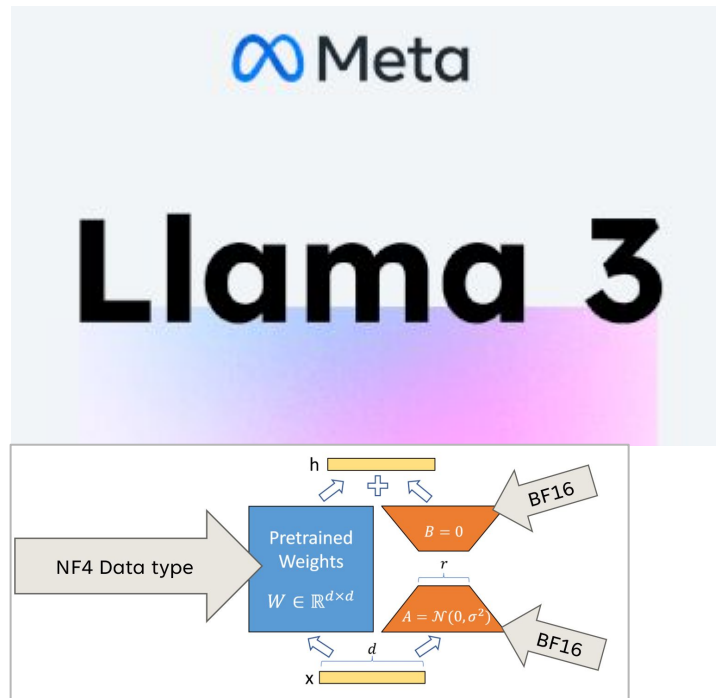
| Model                       | Strengths for Local RAG                                                                                                                             | Limitations                                                                               | Best Use Case                                              |
|-----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|------------------------------------------------------------|
| Phi-4 (MosaicML)            | Excellent instruction-following; strong generation accuracy; available in 1.3B, 6B, 12B parameters; memory-efficient; supports local GPU deployment | Larger variants (6B, 12B) require high-end GPUs for fast inference                        | Local QA, document summarization, secure RAG setups        |
| LLaMA 3                     | High-quality language understanding; good community support; optimized for instruction tuning                                                       | Larger models (13B+) need significant VRAM; smaller models less accurate on complex tasks | Research, prototyping, RAG with vector databases           |
| MPT (MosaicML's MPT series) | Open weights; efficient on consumer GPUs; fast generation; strong instruction-following                                                             | Slightly less accurate on specialized domains compared to Phi-4                           | RAG for general knowledge or enterprise internal documents |

# Methodology [3/6]

**Objective 1: Design and configure three human-AI content-generation workflows**

## B. Fine-tuned LLM using Zambia-specific corpus

- Firstly, Zambian data was collected for fine tuning Llama3
- Secondly, compared the two fine tuning techniques
  - Parameter Efficient Fine Tuning and Full Fine-Tuning
  - Fine tuned Llama3 using Quantized Low Rank Adaptation (QLoRA)
- Thirdly, Publicly deployed the Fine tuned LLM on HuggingFace

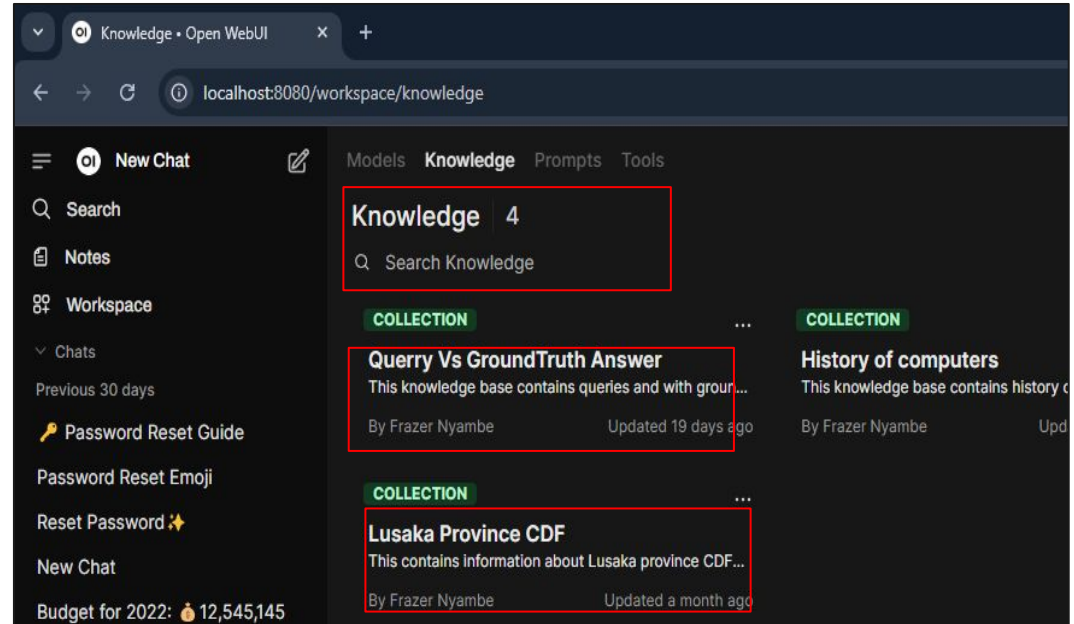


# Methodology [4/6]

## Objective 1: Design and configure three human-AI content-generation workflows

### C. Hybrid RAG pipeline using curated Zambia dataset

- ❏ Open WebUI RAG prototype successfully implemented with a knowledge base
- ❏ This enabled the storage of information as vector embeddings to reduced hallucinations and improve response accuracy when from a pre-trained LLM.



# Methodology [5/6]

## Objective 2: Benchmark factual accuracy, text grounding and contextual recall

The following Metrics will be used to systematically evaluate the three content generation systems using identical prompts and Zambia-related test topics

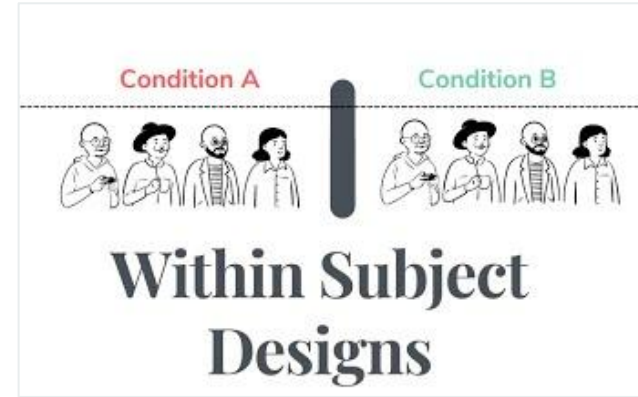
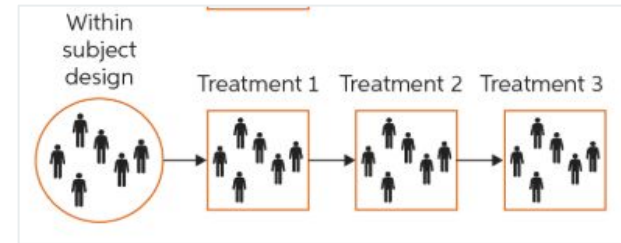
- ❑ **Factual accuracy:**
  - ❑ Measures correctness of generated information
- ❑ **Faithfulness/groundedness:**
  - ❑ Determines whether generated statements are supported by retrieved/reference information
- ❑ **Contextual recall:**
  - ❑ Measures how much relevant information is incorporated



# Methodology [6/6]

## Objective 3: Evaluate the effect of the optimized platform on student editors

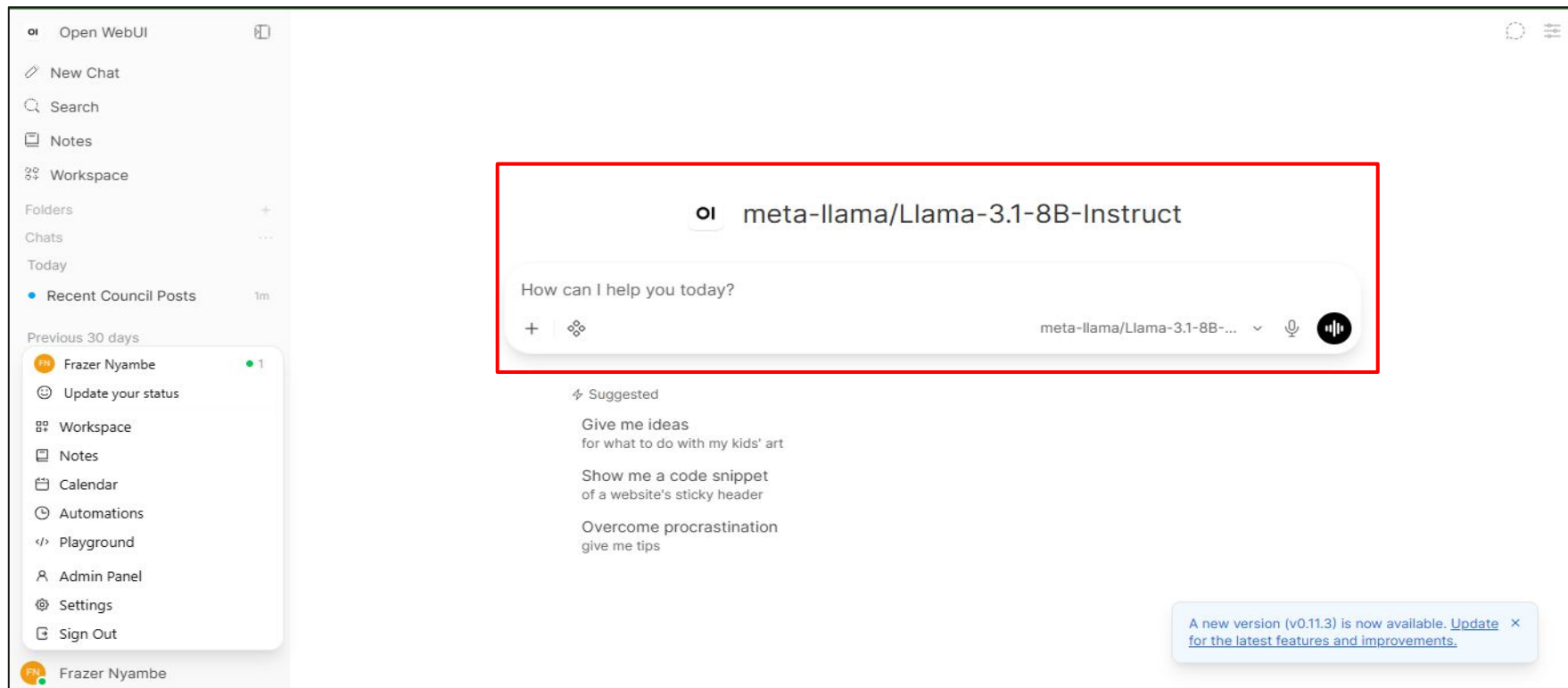
- ❑ A Controlled experiment will be conducted to evaluate the effect of the optimised platform on student editors
- ❑ **Approach:** Within subject study
- ❑ **Participants:** Computer Science students
- ❑ **Task design:**
  - ❑ Participants will be given a task to edit wikipedia article using the three software tools (Base open LLM, fine tuned LLM and RAG system)
  - ❑ Participants will then be given a questionnaire to fill out based on their experimental task



# Preliminary Findings and Discussion [1/7]

## Objective 1: Design and configure three human-AI content-generation workflows

A. **Base open-source LLM:** A software tool was successfully implemented with a pre-trained Llama3 without Zambia specific fine tune or external retrieval



## Preliminary Findings and Discussion [2/7]

## Objective 1: Design and configure three human-AI content-generation workflows

## B. Fine-tuned LLM using Zambia-specific corpus

**Results:** Fine tuning data was successfully extracted from the Local councils and Universities.

```
# Iterate through the list of specific API endpoints and extract posts
for url in SPECIFIC_API_ENDPOINTS:
    extract_and_save_posts_from_url(url)

print("\n--- All specific API endpoint extractions complete ---")

...

Connecting to: https://www.lcc.gov.zm/?rest_route=/wp/v2/posts...
/usr/local/lib/python3.13/dist-packages/urllib3/connectionpool.py:1097: Inse
warnings.warn(
Successfully retrieved 10 posts from lcc.
Title: GOVERNMENT UPSKILLS IT AND PR TEAMS IN WEBSITE MANAGEMENT AND CYBER
Date: 2025-11-28T14:49:53
Link: https://www.lcc.gov.zm/government-upskills-it-and-pr-teams-in-websi
- - -
Title: MAYOR CHITANGALA TOURS ONGOING CHILENJE MARKET REHABILITATION WORKS
Date: 2025-11-26T13:47:30
Link: https://www.lcc.gov.zm/mayor-chitangala-tours-ongoing-chilenje-mar
- - -
Title: CHINIKA WARD 11 COUNCILLOR WRITES
Date: 2025-11-26T10:36:00
Link: https://www.lcc.gov.zm/chinika-ward-11-councillor-writes/
```

```

tokenized_dataset = dataset.map(tokenize_function, batched=True, fn_kwargs={
    print("Dataset tokenized successfully.")
    print(tokenized_dataset)
    # The dataset needs to be formatted for the model (e.g., removing original text)
    # For CausalLM, we often just need 'input_ids' and 'attention_mask'
    # Ensure 'labels' is not removed if you intend to use it for training
    tokenized_dataset = tokenized_dataset.remove_columns([col for col in df_filter
    print("Cleaned tokenized dataset columns:")
    print(tokenized_dataset.column_names)

... Map: 100% ██████████ 4131/4131 [00:00<00:00, 11770.52 examples/s]
Dataset tokenized successfully.
Dataset({
  features: ['university', 'province', 'source_url', 'record_id', 'record_type', 'text']
  num_rows: 4131
})
Cleaned tokenized dataset columns:
['input_ids', 'attention_mask']

```

## Preliminary Findings and Discussion [3/7]

## Objective 1: Design and configure three human-AI content-generation workflows

## B. Fine-tuned LLM using Zambia-specific corpus

**Results:** After applying QLoRA, only 13 million parameters were modified out of 8 billion parameters without retraining the entire network.

```
# 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...
```

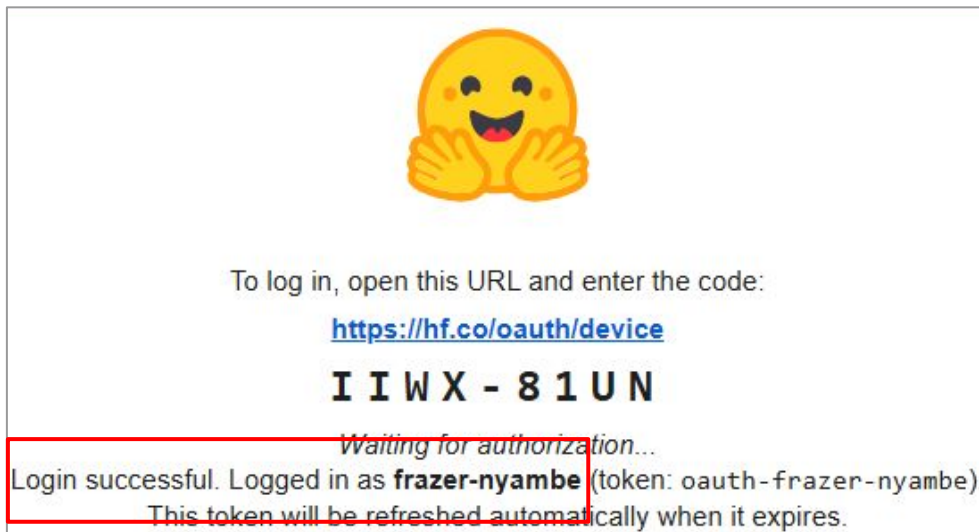
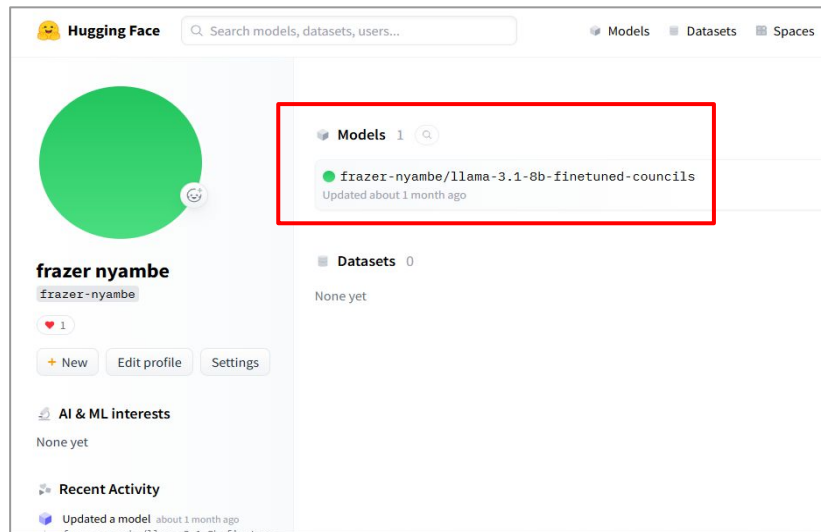
```
config.json: 100% ██████████ 826/826 [0]
tokenizer_config.json: 100% ██████████ 5
tokenizer.json: 100% ██████████ 9.09M/9.
special_tokens_map.json: 100% ██████████
model.safetensors.index.json: 100% ██████████
Download complete: : ████████ 13.8GB, 195MB/s
Reconstruction complete: 100% ████████ 16.1GB / 16.1GB, 90.3MB/s
Fetching 4 files: 100% ██████████ 4/4 [0]
Loading weights: 78% ██████████ 227/2
```

# Preliminary Findings and Discussion [4/7]

**Objective 1: Design and configure three human-AI content-generation workflows**

## B. Fine-tuned LLM using Zambia-specific corpus

**Results:** The Fine Tuned Llama 3 was successfully deployed on Hugging Face.

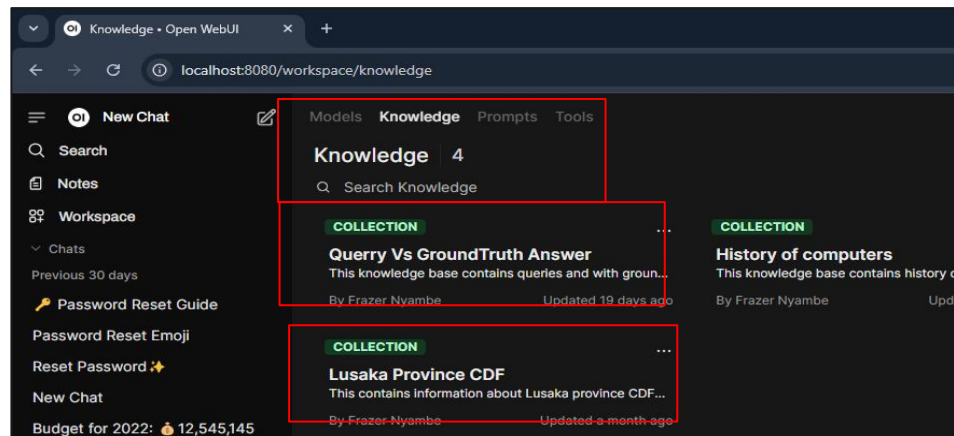
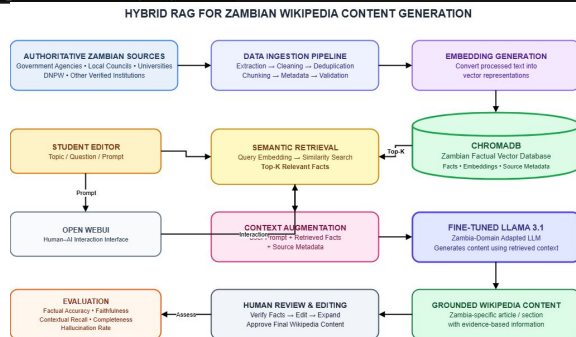


# Preliminary Findings and Discussion [5/7]

## Objective 1: Design and configure three human-AI content-generation workflows

### C. Hybrid RAG pipeline using curated Zambia dataset

- ❑ The Open WebUI RAG prototype successfully demonstrated localized document ingestion and context-aware generation
- ❑ **Results:** Optimized chunking and vector search reduced hallucinations and improved response accuracy when using internal data.



# Findings and Discussion [6/7]

## Objective 2: Benchmark factual accuracy, text grounding and contextual recall



### Anticipated outcomes:

- Controlled generation using identical prompts and Zambia-related test topics

| Model          | Factual accuracy | Text grounding | Contextual recall |
|----------------|------------------|----------------|-------------------|
| Base LLM       | ✓                | ✓              | X                 |
| Fine Tuned LLM | ✓                | ✓              | X                 |
| Hybrid RAG     | ✓                | ✓              | ✓                 |

# Findings and Discussion [7/7]

## Objective 3: Evaluate the effect of the optimized platform on student editors

### Anticipated outcomes:

- Controlled generation using human participants on the three systems

| Evaluation metric        | Base LLM | Fine-tuned LLM | Hybrid RAG | Measurement                                               |
|--------------------------|----------|----------------|------------|-----------------------------------------------------------|
| Task completion time ↓   | ✓        | ✓              | ✓          | Time taken to complete the assigned editing task          |
| Words produced ↑         | ✓        | ✓              | ✓          | Total words produced within the task                      |
| Content expansion rate ↑ | ✓        | ✓              | ✓          | Increase in article/content size after using the workflow |

# Conclusion

- ❑ The study developed and configured three human-AI content-generation workflows: a base LLM, a fine-tuned LLM, and a Hybrid RAG system. Initial implementation confirmed that curated Zambian datasets can be integrated effectively to facilitate localized content generation.
- ❑ Initial findings show that the Hybrid RAG approach improves factual grounding and minimizes hallucinations compared with the standalone LLM. This demonstrates its potential to generate more accurate and evidence-supported Wikipedia-style content on Zambia.
- ❑ Preliminary findings indicate that the human-AI platform can enhance the efficiency and depth of content creation by providing students with accessible, verified information. Further user evaluation will assess its impact on editing efficiency, content completeness, and overall user experience.

# References



- [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.
- [2] Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L. and Shazeer, N. 2018. Generating Wikipedia by Summarizing Long Sequences.
- [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.

# END OF PRESENTATION

