

# BEDRU YIMAM AHMED

Lecturer in Information Technology | Multilingual NLP & Speech Researcher

Kombolcha, Ethiopia · +251 99 630 6284 · [bedruy4@gmail.com](mailto:bedruy4@gmail.com)  
GitHub: [github.com/bedr-ux](https://github.com/bedr-ux) · Google Scholar: [Bedru Yimam Ahmed](#)

## PROFESSIONAL SUMMARY

Multilingual NLP, Speech researcher and Lecturer in Information Technology at Wollo University (KIOT, Ethiopia) with 5 years of university teaching and undergraduate project supervision. My research focuses on low-resource foundation models, speech recognition, and grounded retrieval systems for Ethiopic-script languages (primarily Amharic). Co-author of peer-reviewed papers in *Scientific Reports* (Nature Portfolio) and the AfricaNLP Workshop (ACL Anthology). Author of public empirical benchmarks evaluating custom subword tokenization algorithms (BPE, WordPiece, Unigram), current foundation LLM tokenizer efficiency and token taxation, parameter-efficient fine-tuning (LoRA) forgetting dynamics, and dense versus sparse retrieval in RAG pipelines. Seeking a Ph.D. position in Computer Science / Artificial Intelligence focused on multilingual language modeling, tokenizer efficiency, and retrieval-augmented generation for underserved African languages.

## RESEARCH INTERESTS

- **Multilingual NLP:** Subword tokenization efficiency, vocabulary expansion/grafting, cross-lingual transfer, and morphology-aware representations for Semitic scripts.
- **Parameter-Efficient Fine-Tuning:** LoRA/QLoRA adaptation, 4-bit quantization (NF4), and measuring cross-lingual catastrophic forgetting under zero-replay constraints.
- **Information Retrieval & RAG:** Dense vs. sparse (BM25) failure modes in non-Latin scripts, reciprocal rank fusion (RRF), and hallucination mitigation in low-resource QA.
- **Low-Resource Speech AI:** Acoustic modeling, Whisper adapter fine-tuning for African languages, and speech-to-speech voice search pipelines.

## EDUCATION

**Master of Science (M.Sc.) in Information Technology** 2019 – 2020  
Bahir Dar Institute of Technology (BiT), Bahir Dar University, Ethiopia Cumulative GPA: 3.72 / 4.00  
· **Thesis:** “Graph-based dependency parsing for Amharic Language using Deep Learning”  
· **Advisor:** Dr. Tesfa Tegegne

**Bachelor of Science (B.Sc.) in Information Technology** 2014 – 2017  
Woldia University, Ethiopia Cumulative GPA: 3.86 / 4.00  
· Graduated with *Very Great Distinction*

## PEER-REVIEWED PUBLICATIONS & PREPRINTS

1. **Multimodal Based Amharic Fake News Detection Using CNN and Attention-Based BiLSTM**  
Alemu Desu Geto, Eshete Derb Emiru, Nuru Seid, **Bedru Yimam Ahmed**, et al. [DOI: Nature Scientific Reports]  
*Scientific Reports* (Nature Portfolio), October 2025.  
· Co-designed a dual-stream neural architecture fusing visual features and contextual word embeddings across 23,856 news articles, achieving 98% test classification accuracy.
2. **Automatic Speech Recognition for African Low-Resource Languages: Challenges and Future Directions**  
Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, **Bedru Yimam Ahmed**, et al. [ACL Anthology]  
*Proceedings of the 6th Workshop on African Natural Language Processing (AfricaNLP 2025)*, ACL Anthology, July 2025.  
· Co-authored an empirical survey detailing dataset scarcity, acoustic variability, and compute bottlenecks across indigenous African language families.
3. **Whispering in Amharic: Fine-Tuning Whisper for Low-Resource Language**  
Dawit Ketema Gete, **Bedru Yimam Ahmed**, Tadesse Destaw Belay, et al. [arXiv:2503.18485]  
*arXiv Preprint*, arXiv:2503.18485 [cs.CL], March 2025.  
· Fine-tuned OpenAI Whisper for Amharic speech across Mozilla Common Voice, FLEURS, and BDU-Speech corpora, significantly reducing Word Error Rate (WER) under noisy acoustic conditions.

## RESEARCH ARTIFACTS & OPEN-SOURCE BENCHMARKS

**Amharic Sub-Word Tokenization Algorithm Analysis (BPE, WordPiece, Unigram)** [2026]  
*Principal Author* [github.com/bedr-ux/Amharic-Sub-word-Tokenization-Algorithm-Analysis](https://github.com/bedr-ux/Amharic-Sub-word-Tokenization-Algorithm-Analysis)  
· Implemented, trained, and comparatively analyzed custom subword tokenizers (Byte-Level BPE, WordPiece, and Unigram) on the bilingual MasakhaNEWS dataset (Amharic and English).

- **Empirical Findings:** Byte-Level BPE and WordPiece achieved virtually identical optimal compression at 16k vocabulary (fertility: 1.399 vs. 1.400 tokens/word; 8.93 bytes/token), significantly outperforming Unigram (1.593 fertility; 7.85 bytes/token).
- Evaluated Ge'ez orthographic normalization (Fidel homophone merging and punctuation isolation, reducing BPE fertility from 1.529 to 1.399), vocabulary scaling ( $4k \rightarrow 32k$ , showing diminishing returns past 16k), and monolingual vs. joint bilingual allocation (fertility shift from 1.399 to 1.639).

### Amharic LLM & RAG Retrieval Benchmark

[2026]

Lead Researcher & System Architect

[github.com/bedr-ux/Amharic-LLM-RAG-Benchmark](https://github.com/bedr-ux/Amharic-LLM-RAG-Benchmark)

- Built an evaluation suite comparing 3 retrievers (Dense MiniLM+FAISS, Ethiopic-aware BM25Okapi, and equal-weight RRF Hybrid) across 3 generators (Qwen2.5-0.5B, Tiny Aya Earth 3.35B 4-bit, and Amharic-LLaMA-0.4B) on 2,617 QA pairs (dagim/amharic-qa).
- **Empirical Findings:** Sparse BM25 strongly outperformed dense retrieval on Amharic (Recall@3: 0.900 vs. 0.117; MRR: 0.836 vs. 0.120). BM25 RAG yielded a  $5.0\times$  Token F1 surge on Tiny Aya Earth ( $0.038 \rightarrow 0.191$ ; SacreBLEU  $0.89 \rightarrow 6.93$ ; paired  $t$ -test  $p = 4.3 \times 10^{-15}$ , surviving Bonferroni correction).

### Amharic Tokenization Benchmark across Current Foundation LLMs

[2026]

Principal Author

[github.com/bedr-ux/Amharic-Tokenization-Benchmark](https://github.com/bedr-ux/Amharic-Tokenization-Benchmark)

- Audited and benchmarked the tokenizer efficiency of current large language models (LLaMA-3-8B, GPT-4o, Tiny Aya Earth, XLM-RoBERTa, AfroXLMR, and specialized checkpoints) on 1,012 parallel sentence pairs from FLORES-200 devtest.
- **Empirical Findings:** Quantified the severe "Token Tax" in current LLMs for Amharic—revealing a  $9.59\times$  fertility disparity in LLaMA-3 vs. English (11.89 vs. 1.24 tokens/word; 0.423 CPT vs. 4.857 in English; 95% bootstrap CI [9.48, 9.70]) and 0.91 tokens/byte, proving catastrophic byte-level fallback that degrades effective context windows and inflates inference costs.

### Amharic LoRA Cross-Lingual Adaptation Study

[2026]

Principal Author

[github.com/bedr-ux/Amharic-LoRA-Experiment](https://github.com/bedr-ux/Amharic-LoRA-Experiment)

- Fine-tuned Qwen2.5-0.5B via LoRA ( $r = 16, \alpha = 32$ , updating 0.44% of weights) on 1.78M tokens of deduplicated Amharic Wikipedia text on a single Tesla T4 GPU.
- **Empirical Findings:** Reduced Amharic BPC by 31.1% ( $6.163 \rightarrow 4.248$ , 95% bootstrap CI  $[-31.4\%, -30.7\%]$ ), while precisely measuring an 18.8% catastrophic forgetting cost on English under zero-replay conditions.

## ACADEMIC TEACHING & MENTORSHIP EXPERIENCE

### Lecturer, Department of Information Technology

Nov 2020 – Present

Wollo University, Kombolcha Institute of Technology (KIOT), Ethiopia

- Taught undergraduate courses: Artificial Intelligence, Data Structures & Algorithms, Advanced Database Systems, Python, Java, C++, and Web Architectures across 10+ academic semesters.
- Primary academic advisor for 80+ senior undergraduate thesis projects, mentoring students through deep learning model implementation, experimental design, and scientific writing.

### Graduate Assistant II / Assistant Lecturer

2018 – 2019

Wollo University, Kombolcha Institute of Technology (KIOT), Ethiopia

- Conducted laboratory instruction and programming coursework for Systems Programming and Database Systems.

## TECHNICAL SKILLS

- **LLM & NLP Tooling:** PyTorch, Hugging Face (transformers, peft, tokenizers, datasets, accelerate), BitsAndBytes (4-bit NF4), LoRA/QLoRA, SentencePiece.
- **Retrieval & Evaluation:** FAISS, ChromaDB, LangChain, BM25Okapi, Reciprocal Rank Fusion (RRF), Token F1, SacreBLEU, Recall@k, MRR, Bits Per Character (BPC), RAGAS framework.
- **Speech Processing:** Whisper, Massively Multilingual Speech (MMS), Word Error Rate (WER).
- **Programming & Environment:** Python, C++, Java, PHP, JavaScript, SQL, Git/GitHub, L<sup>A</sup>T<sub>E</sub>X.

## ACADEMIC SERVICE & REFEREES

**Service:** Reviewer for Wollo University Internal Thematic Research Grants (2026); Technical Reviewer for the 6th National Annual Research Conference on Science, Technology, and Innovation (ASTISD-2024).

### Referees:

- **Dr. Tesfa Tegegne** – Associate Professor, Bahir Dar University · [tesfat@gmail.com](mailto:tesfat@gmail.com) (M.Sc. Advisor)
- **Dr. Zebiba Ali** - Assistant Professor & AI Research Director Directorate, Wollo University · [alizebiba@gmail.com](mailto:alizebiba@gmail.com)