KELECHI EZEMA

kelechiezema.com | kelechi.ezema@colorado.edu | linkedin.com/in/kelechi-ezema
Boulder, Colorado - 80302, USA

EDUCATION

- **PhD in Computer Science** *Aug 2023 - Current*
  University of Colorado Boulder, USA
  Advisor: Prof. Sidney DMello
- **MSc Computer Science -** *Dec. 2025*
  University of Colorado Boulder, USA
  Advisor: Prof. Sidney DMello
- **Bachelor of Science (B.Sc), Computer Science with Statistic** *June 2021*
  University of Nigeria, Nsukka (UNN), Nigeria
  Supervisor: Dr. Blessing Ogbuokiri

RESEARCH INTEREST

Educational NLP, Responsible AI, Speech Processing, Human-Computer Interaction (HCI), Multimodal LLM Evaluation, Agentic AI Evaluation

PROFESSIONAL/WORK EXPERIENCE

Applied Science Intern. *May 2026 - July 2026*
Microsoft Corporation
- Led LLM evaluation efforts for AI-generated summaries on Microsoft OneNote app and developed rigorous methodology to measure LLM summaries, also engineered an end-to-end benchmarking system for Copilot Notebook agentic workflows.

Graduate Research Assistant *Aug. 2023 - Current*
Emotive Computing Lab - CU Boulder
- Led three research projects as an NLP and HCI researcher for an NSF and LEVI-funded project aimed at enhancing human tutoring through AI, with a focus on improving math scores for low-income students.

Summer Intern *June 2025 - Aug. 2025*
Rising Academies (Rori and Tari Chatbots Products)
- Led the technical evaluation and user study assessment of LLM-powered educational chatbots (accessed via WhatsApp), providing an accessible AI-powered assistant for teachers and students in low-income households.

Graduate Teaching Assistant *Fall '23, '24 & Spring 25*
Computer Science Department - CU Boulder
- Mentored 35 undergraduates across two C++ and software development recitations and spearheaded final project design and rubrics for over 800 students.

SELECTED RESEARCH PROJECT

WorkSpace Bench: A product-wide Evaluation of Copilot Notebook (CPNB) AI quality
Microsoft Corporation
- Developed an end-to-end benchmarking framework to evaluate the agentic capabilities of Copilot Notebook across multi-step workflows, including Q&A, summarization, artifact generation, and reference editing.
- Partnered with two external vendors to curate over 700 high-coverage queries and golden response across three distinct domains, ensuring comprehensive test coverage across all supported file types, user queries, and grounding documents.
- Designed a multi-tiered evaluation methodology combining standardized external benchmarks, user quality rubrics, and engineered assertions to reliably detect model regressions and measure performance.

Impact of Racial Bias in Automatic Speech Recognition (ASR) on Downstream LLM Classifiers
Emotive Computing Lab - CU Boulder
- Investigated the performance disparities of ASR models across different racial demographics, highlighting biases in speech recognition systems with over 23% difference in ASR performance across investigated model
- Analyzed the downstream effects of ASR biases on *RoBERTa* classifiers, specifically in quantifying tutor performance in real-world settings.
- Developed and implemented strategies to mitigate bias in ASR models, improving model accuracy and fairness across various racial groups, contributing to more equitable speech recognition systems.
- The research received Honorable Mention at the prestigious CHI Conference, recognizing its impact on responsible AI development.

Automatic measurement and prediction of student engagement through speech
Emotive Computing Lab - CU Boulder
- Developed NLP tools that leverage children's speech to predict cognitive engagement and in real-world tutoring programs achieving significant performance after controlling over other competing variables
- Conducted a study on data augmentation to improve the predictive accuracy of the RoBERTa encoder model when applied to noisy classroom environments using children's speech data.

- Evaluated the trade-offs between encoder models (RoBERTa) and generative models (GPT) in educational applications, focusing on how children's speech influences model performance in real-world settings.

LLM Evaluation (Rori and Tari Chatbots Products)

Rising Academies

- Led the evaluation and user engagement assessment of LLM-powered educational chatbots used by teachers and students in Sub-Saharan Africa, resulting in a 15% increase in user engagement through the integration of multimodal features such as speech and image recognition.
- Developed a comprehensive evaluation framework to analyze user conversation logs across multiple turns, ensuring alignment with educational goals.
- Applied advanced prompt engineering techniques to address user engagement gaps in Anthropic's Claude-powered chatbots, significantly improving user interaction and content relevance.

Quality of Tutoring Framework (QTF): Using ML to Augment Human Tutor Performance

Emotive Computing Lab - CU Boulder

- Developed statistical models to identify correlations between different evaluation frameworks used to assess tutor performance, leveraging ML techniques to improve math tutoring effectiveness for low-income students across several school districts in the USA.
- Integrated natural language processing (NLP) tools to process tutors' and students' utterances in a tutoring system, and applied LLM fine-tuning for better contextual learning and feedback.
- Collaborate with cross-functional teams to identify key learning dimensions to improve tutor effectiveness.

N-gram based de-baising strategy for Automatic Speech Recognition (ASR)

Emotive Computing Lab - CU Boulder

- Proposed a novel N-gram data augmentation strategy, extracting short-phrase utterances across diverse racial demographics to mitigate bias in Automatic Speech Recognition (ASR) systems.
- Spearheaded the fine-tuning of a CTC-based Wav2Vec model, combining N-gram datasets with a counterfactual learning objective to improve transcription equity.
- Achieved a 13% overall performance gain and reduced the demographic performance gap by 17% compared to baseline models and conventional fine-tuning methods.
- Validated the approach's generalizability across the FairSpeech benchmark and the LEVI educational corpus, successfully evaluating Black and White speaker cohorts; paper currently under review.

PUBLICATIONS AND TALK

C=CONFERENCE, P=POSTER/PRESENTATION, J=JOURNALS , T=TALKU=UNDER REVIEW

[C.1] **Kelechi Ezema**, Chelsea Chandler, Rosy Southwell, Niranjan Cholendiran and Sidney D'Mello. (2025) "It feels like we're not meeting the criteria": Examining and Mitigating the Cascading Effects of Bias in Automatic Speech Recognition in Automatic Feedback Interfaces. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 19* [Honorable Mention - Best Paper Award, top 5% across 5k submissions]

[J.1] **Ezema Henry Kelechi**, Blessing Ogbuokiri. Online Result Processing System for Secondary Schools *International Journal of Computer Sciences and Engineering (IJSCE) , Vol.10, Issue.9*

[P.1] **Kelechi Ezema**, Chelsea Chandler, Rosy Southwell, Niranjan Cholendiran and Sidney D'Mello *Examining and Mitigating the Cascading Effects of Bias in Automatic Speech Recognition in Automatic Feedback Interfaces.* at The Winter Text and Discourse Conference (WT&D '25), February 2025, Boulder Colorado, USA.

[U.1] **Kelechi Ezema**, Ekta Sood, Rosy Southwell, Sidney D'Mello (Under Review) Fairness Beyond Benchmark: Evaluating N-gram Based De-baising Strategies for Automatic Speech Recognition.

[U.2] **Kelechi Ezema**, Chinecherem Ezema, Chibuzor Okocha (Under Review) Is Newer Always Better? Evaluating Successive Speaker Diarization Systems Across Conversational Domains.

[C.2] Naome A Etori, **Kelechi Ezema**, Nathaniel Romney Robinson, Davis David, Alfred Malengo Kondoro, Elisha Ondieki Makori, Michael Samwel Mollel, Maria Gini. AfriMMT-EA: Multi-domain Machine Translation for Low-Resource East African Languages. *European Chapter of the Association for Computational Linguistics (EACL '26)*

[U.3] Chelsea Chandler, **Kelechi Ezema**, Sidney K. D'Mello (Under Review) Automated Analysis of Student Discourse as a Measure of Cognitive Engagement in Small Group Tutoring.

[U.4] Tochukwu Eze, **Kelechi Ezema**, Almene De Meran Meguimtsop and Cornelius Adejoro (Under Review) Understanding Privacy in Context: A Systematic Review of Risks, Protections, and Human-Centered Challenges in Multimodal Learning Analytics.

[U.6] Chibuzor Okocha, **Kelechi Ezema**, Christian Grant (Under Review) A Comparative Analysis of Speaker Diarization Models: Evaluating a Dataset for African Accented Speech.

[P.2] Dr. Christina Alston and **Kelechi Ezema** *Innovating the Next Generation of STEM: Workshop presentation* at The Leadership Alliance (LANS) conference held in July 2024 at Hartford, Connecticut, USA.

[C.3] Dorothea French, Robert Moulder, **Kelechi Ezema**, Katharina von der Wense, Sidney K. DMello Linguistic Alignment Predicts Learning in Small Group Tutoring Sessions. Main finding at The 2025 Conference on *Empirical Methods in Natural Language Processing* (EMNLP '25)

[C.4] Tobi Olatunji, Busayo Awobade, Mardhiyah Sanni, Tassallah Abdullahi, Chibuzor Okocha, **Kelechi Ezema**, Devendra Deepak Kayande, Lukman Enegi Ismaila, Gloria Ashiya Katuka. AfriVox: Probing Multilingual and Accent robustness of Speech LLMs. Submitted to *European Chapter of the Association for Computational Linguistics (EACL '26)*

[U.5] Chelsea Chandler, Sandra Sawaya, Jennifer Jacobs, Halley Bowman, **Kelechi Ezema** and Sidney D'Mello (Under Review) Aligning Researcher and Practitioner Theoretical Frameworks in Learning Analytics Applications: A Generalizable Approach and a Case Study.

[U.6] Naome A Etori, Jingyan Wang, Macton Mgonzo, Kevin Obote, Suad Abdulkadir, **Kelechi Ezema**, Michael Samwel Mollel, Maria Gini (Under Review) The Illusion of Fluency: Evaluating Figurative and Cultural Understanding in Multilingual LLMs with AFRI-PROVERB.

[C.4] Mardiyyah Oduwole, Tassallah Abdullahi, Tobi Olatunji, Gloria Ashiya Katuka, Macton Mgonzo, Chibuzor Okocha, Terry Oko-odion, **Kelechi Ezema**, Almene De Meran Meguimtsop, Lukman E. Ismaila. Audio Penalty: Evaluating Clinical Safety of Multimodal LLMs Across Text and Speech in Low-Resource Languages. *Africa at NeurIPS 2026*

Technical Skills

- **Data Analysis & Visualization:** R (lme4, tidyverse, ggplot2), pandas, Tableau and Microsoft PowerBI.
- **Programming & Tools:** Python, SQL, JavaScript, C++, HTML, CSS, React, Laravel, GitHub, and Jupyter Notebook
- **Machine Learning:** PyTorch, TensorFlow, Scikit-learn, Wandb, HuggingFace, LLM fine-tuning, prompt engineering.
- **Speech Recognition:** audioSegmentation, Transformers, librosa, nltk, jiwer, and Levenshtein.

Leadership/Volunteer Experience

- **Program Committee (PC) Member** *March 2026*
  International Conference on Artificial Intelligence in Education (AIED 2026)
  - Served as a peer reviewer in reviewing six full paper academic submissions while engaging in detailed conversation with peer reviewers on merit and demerit of submissions.
  - Reviewed 10 short paper tracks submitted on education advancement with NLP and AI systems while pointing authors to methodology flaws and areas of improvements

- **President** *April. 2025 - Date*
  African Graduate Student Association (AGSA), University of Colorado Boulder
  - Coordinated institution-wide initiatives with over 250 attendees across four universities campuses within Colorado and Wyoming , aimed at enhancing student engagement and fostering a sense of community among graduate students
  - Secured over $5,000 in funding to support the association's programs, directly contributing to the successful execution of events and activities

- **Local Conference Organizer** *Feb. 2025*
  Winter Text and Discourse conference (WT&D '25) held in Boulder, Colorado, USA
  - Led logistical coordination for a nationwide conference held in Boulder, Colorado, ensuring smooth operations and seamless execution of conference activities
  - Designed and produced conference and tutor guides, significantly enhancing the attendee experience and ensuring clear communication throughout the event

- **Graduate Mentor and Research Presentation Judge** *Summer 2024*
  Office of Graduate Access and Retention, University of Colorado Boulder
  - Mentored rising juniors and seniors in the 10-week SMART program, providing guidance and offering innovative insights to emerging researchers
  - Supported students in developing their research project focusing on enhancing their research skills, critical thinking, and enhancing their academic growth.

Honors and Awards

- **NSF Summit for AI Institutes Leadership (SAIL) Travel Grant** *Aug. 2025*
  NSF AI Institutes Virtual Organization (NSF AIVO)
- **Honorable Mention Recognition - Top 5% of 5,000 paper submissions** *March 2025*
  Human Computer Interaction (CHI'25)
- **Research Fellowship - $9,500** *June 2025*
  Handshake Model Validation Expert (MOVE) fellowship [Declined]
- **Dean's and Departmental Excellence Fellowship** *Aug. 2023*
  Computer Science Department, University of Colorado Boulder