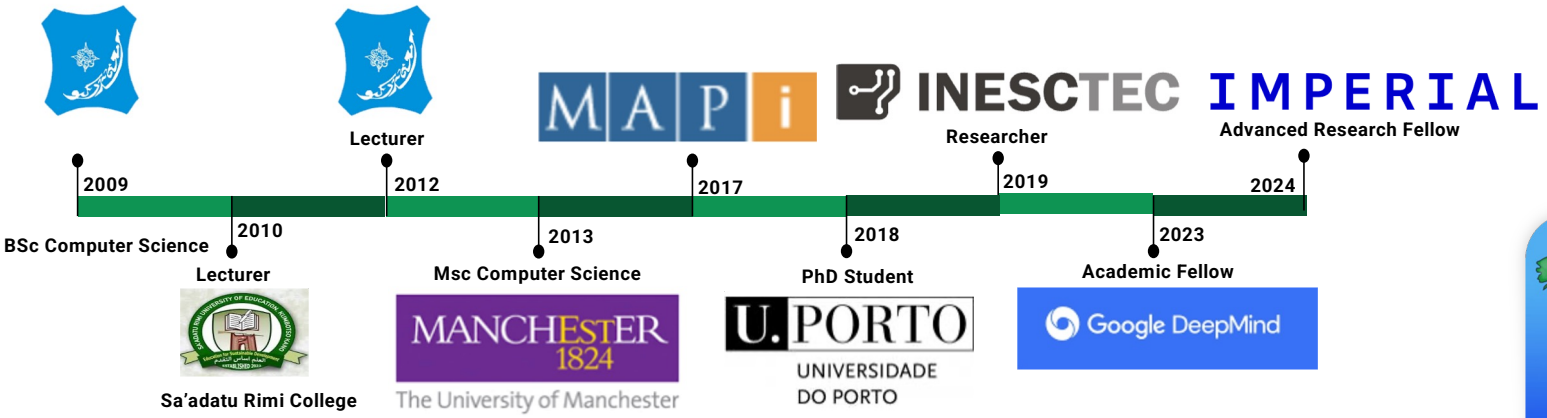
The Illusion of Inclusion

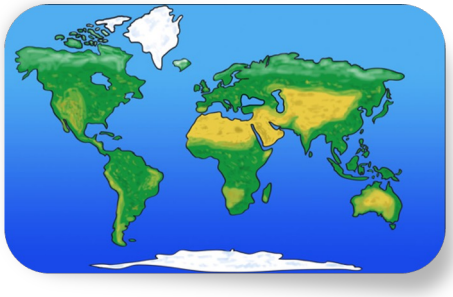
How LLMs Misrepresent African Languages and Cultural Contexts

Shamsuddeen Muhammad
Google DeepMind Academic Fellow,
Advanced Research Fellow,
Imperial College London
Bayero University, Kano

https://shmuhammadd.github.io/

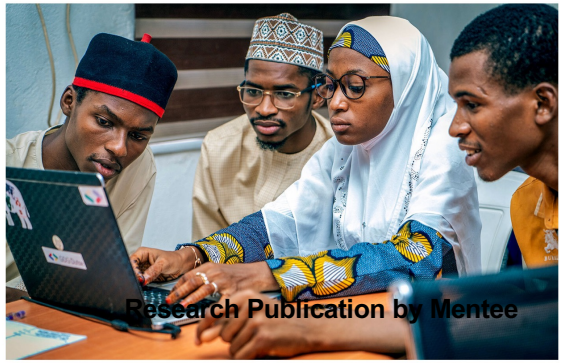
Career Timeline





Mentorship and Capacity Building

AIMS Africa	Co-Founder	Co-Founder
Cameroon and **South Africa**	Arewa Data Science Academy	HausaNLP



In May, fellows of the Arewa Data Science Academy, a free training program for Nigerian youth who want to learn data science and machine learning, participated in an artificial intelligence hackathon in Nigeria. AREWA DATA SCIENCE

Outline

A Two-Decade Journey of African NLP

The Illusion of Inclusion

Selected Works

AI Research in the Global South

Language Technology in Low-Resource Languages

Significant improvements in language models have primarily benefited English texts

ChatGPT Claude

Mistral Gemini

LLaMA Command R

AI often mangles African languages. Local scientists and volunteers are taking it back to school

A network of thousands of coders and researchers is working to develop translation tools that understand their native languages

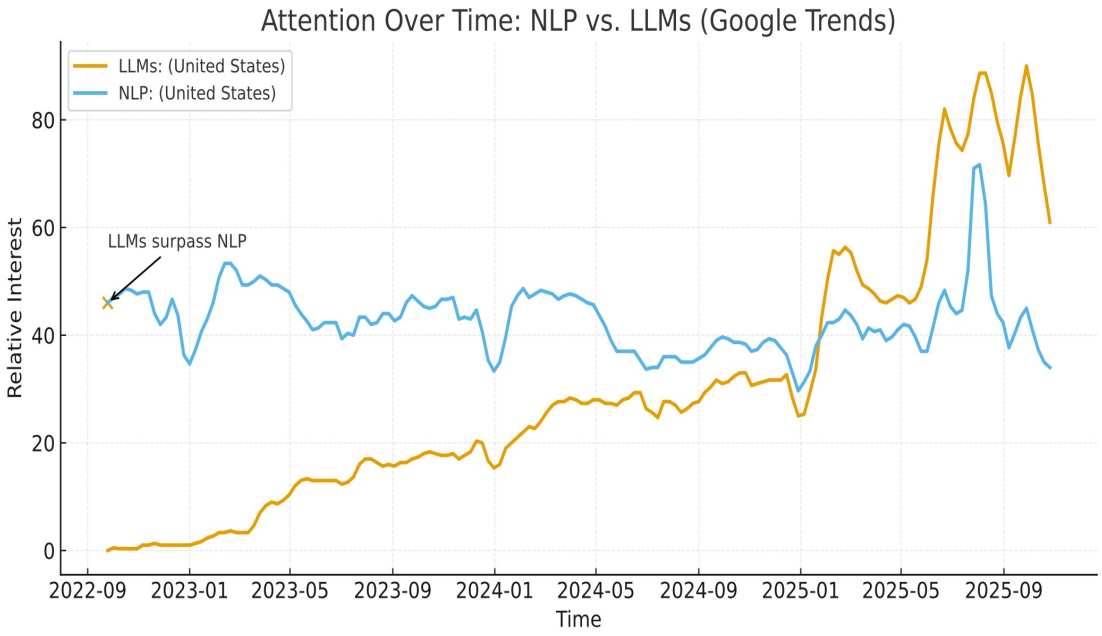
20 JUL 2023 · 7:00 AM · BY SANDEEP RAVINDRAN

https://www.science.org.

LLMs Are Surging - But African Language Are Left Behind

Growth in AI, not Inclusion

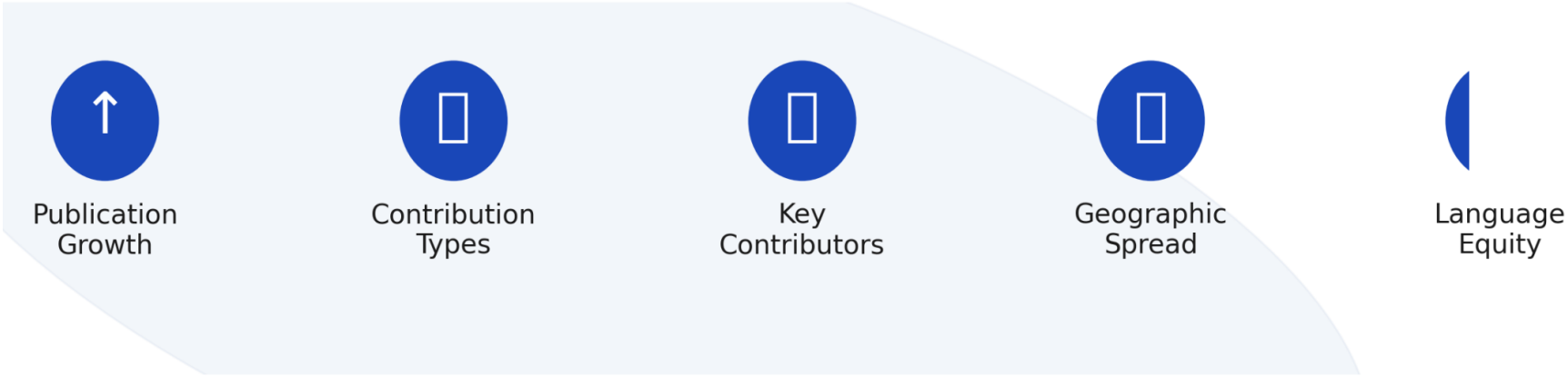
- LLMs is redefining NLP research priorities
- 60%+ of ACL 2025 submissions mention “LLM”
- Africa’s > 2,000 languages → **minimal inclusion**
- We cannot fix what we do not measure**, where we are**?



A Two-Decade Journey of African NLP

The Rise of African NLP: Contributions, Contributors, and Community Impact (2005–2025) Belay, Tadesse et al. 2025.

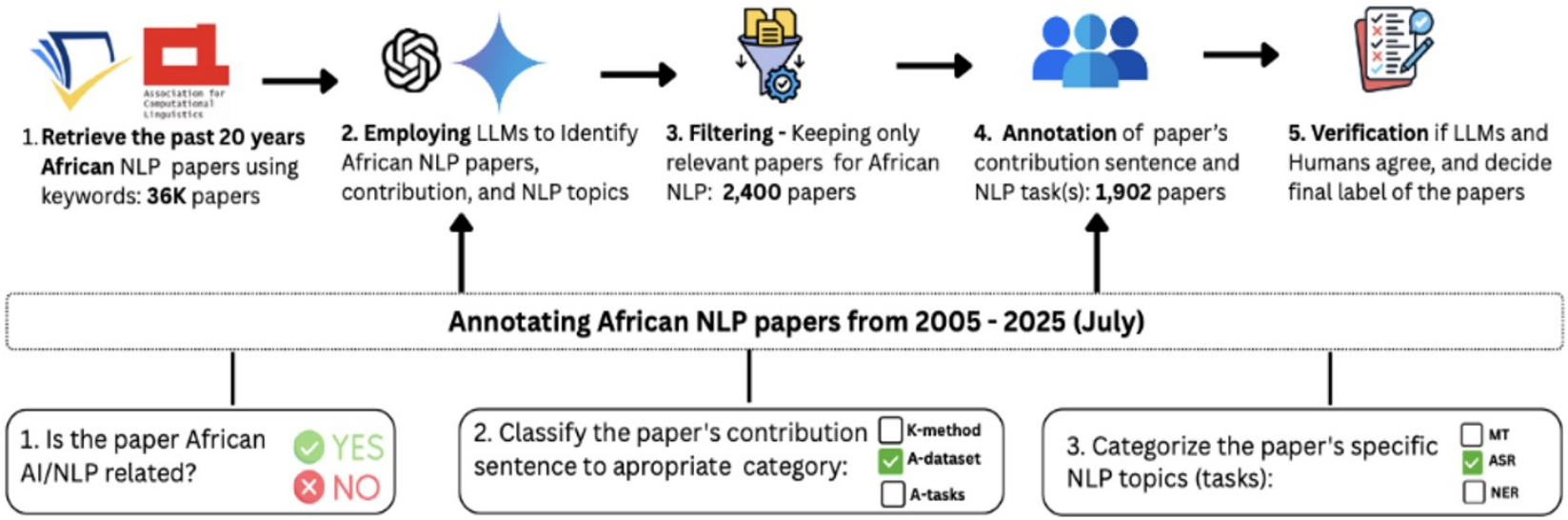
How We Measure Progress in AfricaNLP



The Rise of African NLP: Contributions, Contributors, and Community Impact (2005–2025) Belay, Tadesse et al. 2025.

The Rise of African NLP (2005-2025)

Data Collection and Annotation

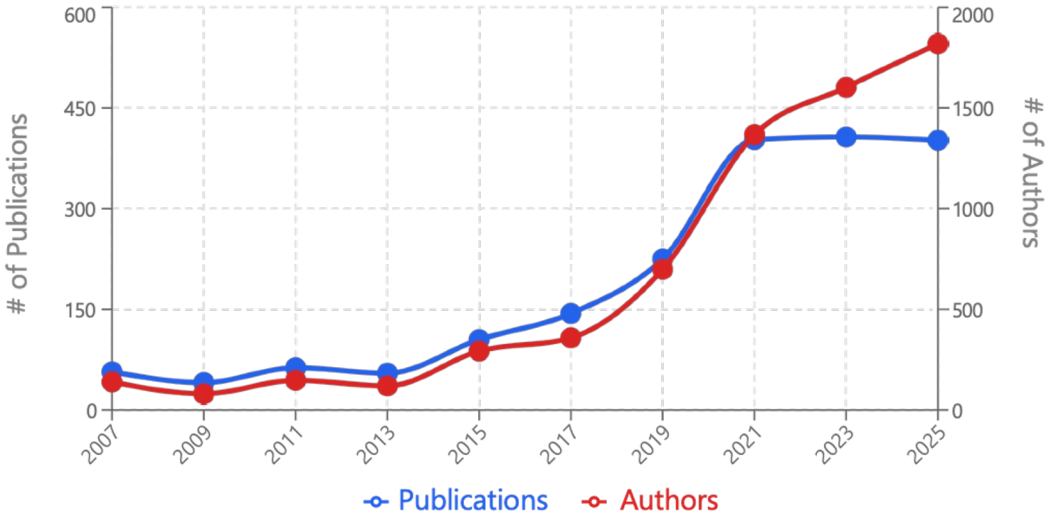


The Rise of African NLP (2005-2025)

Two-Decade Analysis of African NLP

'The limits of my language mean the limits of my world.'…. Wittgenstein

The languages we build models for *determine who participates in the digital world.*



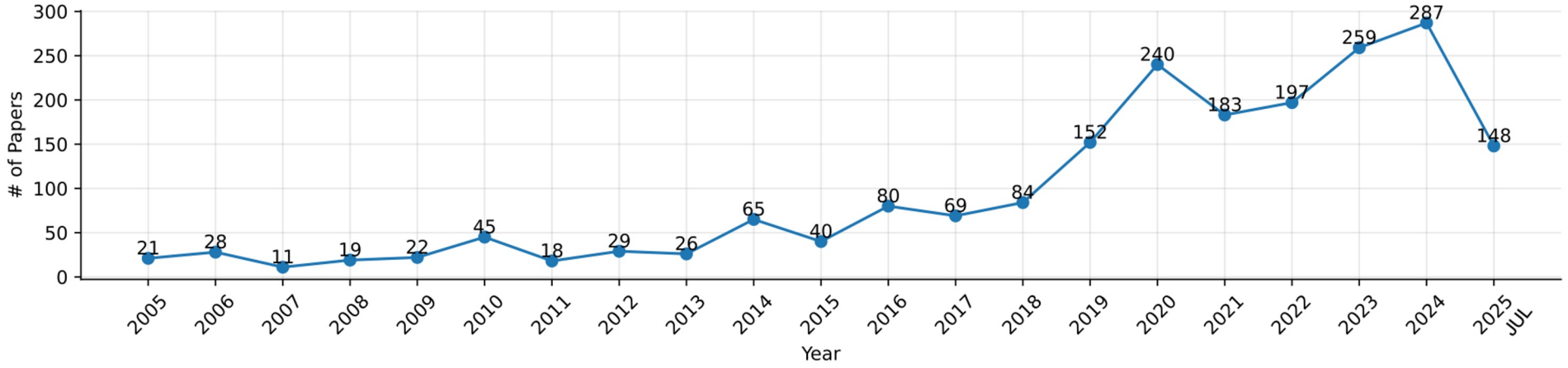
From a few dozen papers to over 400 publications and nearly 2,000 contributors

The Rise of African NLP: Contributions, Contributors, and Community Impact (2005–2025) Belay, Tadesse et al. 2025.

The Rise of African NLP (2005-2025)

Two-Decade Analysis of African NLP

Papers from 1.9K AfricaNLP papers spanning 2005–2025.



Sharp increase from 2019 onward, reflecting the rise of community-driven initiatives (e.g., *Masakhane*, and *AfricaNLP Workshops*).

State of AI in Africa

Community Driven Participatory research

Compared to other regions where the AI ecosystem is shaped by Universities, big corporations or strong policies and regulation frameworks:

Africa's AI ecosystem is dominated by grassroots movements, such as 'Deep Learning Indaba' and 'Data Science Africa'.

AI and the Future of Work in Africa White Paper O'Neill, J., et al. (2024, June).

AfricaNLP Contributors

Top Institutions and Funders Driving AfricaNLP (2005–2025)

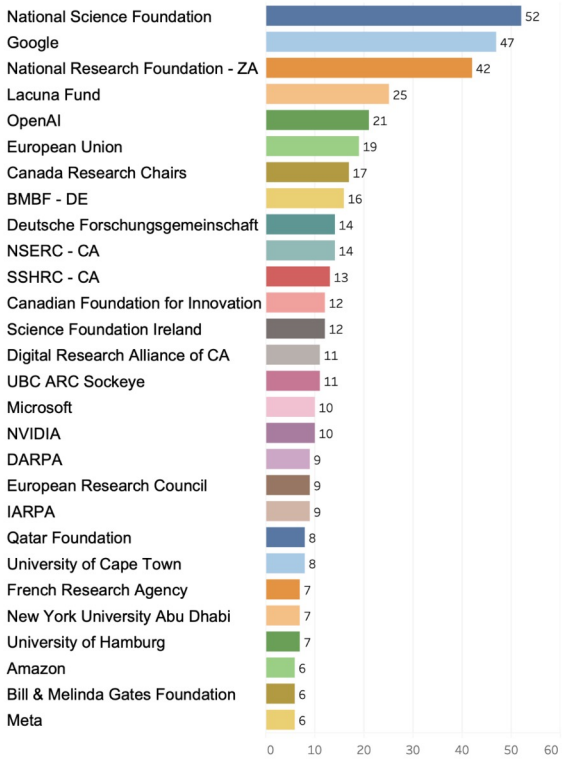
*What are the top organizations supporting the development of AfricaNLP based on evidence from **1.9K NLP papers**?*

Acknowledgments

Shamsuddeen Muhammad acknowledges the support of Google DeepMind and Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre. The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, its funders, or Meridian Institute.

BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages

Extracted from 1.9K NLP papers manually.



AfricaNLP Contributors

Top Institutions and Funders Driving AfricaNLP (2005–2025)

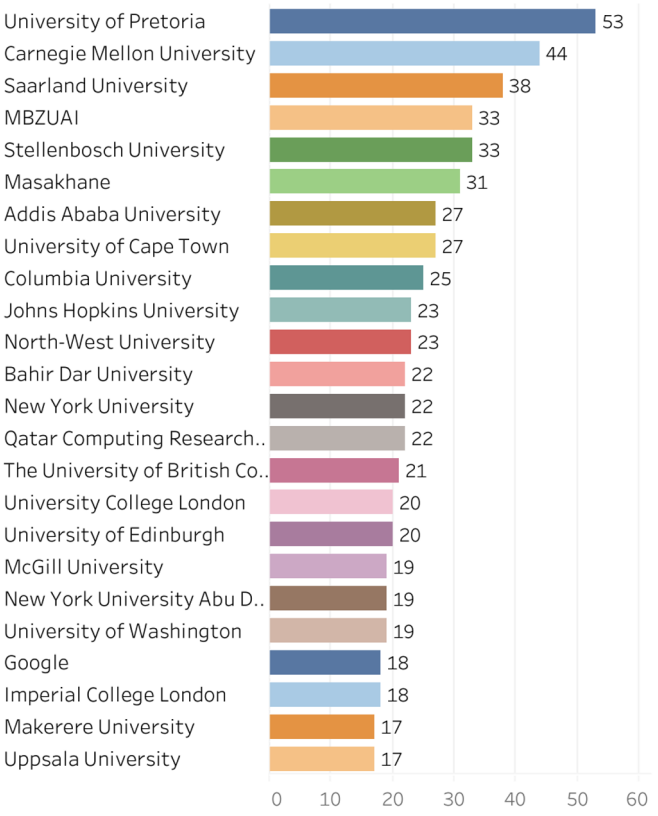
What are the top affiliated institutions, organizations of the authors?

BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages

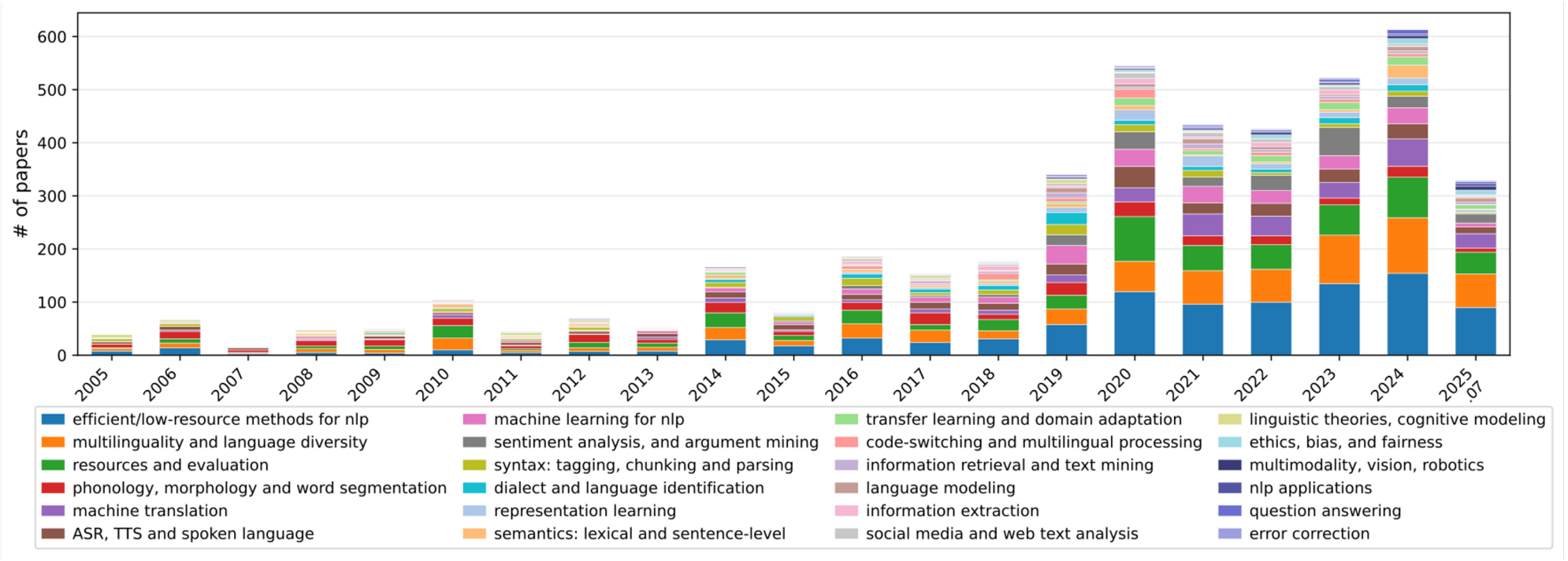
Shamsuddeen Hassan Muhammad[1,2*], Nedjma Ousidhoum[3*], Idris Abdulmumin[4], Jan Philip Wahle[5], Terry Ruas[5], Meriem Beloucif[6], Christine de Kock[7], Nirmal Surange[8], Daniela Teodorescu[9], Ibrahim Said Ahmad[10], David Ifeoluwa Adelani[11,12,13], Alham Fikri Aji[14], Felermino D. M. A. Ali[15], Ilseyar Alimova[31], Vladimir Araujo[16], Nikolay Babakov[17], Naomi Baes[7], Ana-Maria Bucur[18,19], Andiswa Bukula[20], Guanqun Cao[21], Rodrigo Tufiño[22], Rendi Chevi[14], Chiamaka Ijeoma Chukwuneke[23], Alexandra Ciobotaru[18], Daryna Dementieva[24], Murja Sani Gadanya[2], Robert Geislinger[25], Bela Gipp[5], Oumaima Hourrane[26], Oana Ignat[27], Falalu Ibrahim Lawan[28], Rooweither Mabuya[20], Rahmad Mahendra[29], Vukosi Marivate[4,30], Alexander Panchenko[31,32], Andrew Piper[12], Charles Henrique Porto Ferreira[33], Vitaly Protasov[32], Samuel Rutunda[34], Manish Shrivastava[8], Aura Cristina Udrea[35], Lilian Diana Awuor Wanzare[36], Sophie Wu[12], Florian Valentin Wunderlich[5], Hanif Muhammad Zhafran[37], Tianhui Zhang[38], Yi Zhou[3], Saif M. Mohammad[39]

[1]Imperial College London, [2]Bayero University Kano, [3]Cardiff University, [4]Data Science for Social Impact, University of Pretoria, [5]University of Göttingen, [6]Uppsala University, [7]University of Melbourne, [8]IIIT Hyderabad, [9]University of Alberta, [10]Northeastern University, [11]MILA, [12]McGill University, [13]Canada CIFAR AI Chair, [14]MBZUAI, [15]LIACC, FEUP, University of Porto, [16]Sailplane AI, [17]University of Santiago de Compostela, [18]University of Bucharest, [19]Universitat Politècnica de València, [20]SADiLaR, [21]University of York, [22]Universidad Politécnica Salesiana, [23]Lancaster University, [24]Technical University of Munich, [25]Hamburg University, [26]Al Akhawayn University, [27]Santa Clara University, [28]Kaduna State University, [29]Universitas Indonesia, [30]Lelapa AI, [31]Skoltech, [32]AIRI, [33]Centro Universitário FEI, [34]Digital Umuganda, [35]National University of Science and Technology Politehnica Bucharest, [36]Maseno University, [37]Institut Teknologi Bandung, [38]University of Liverpool, [39]National Research Council Canada

Contact: s.muhammad@imperial.ac.uk, OusidhoumN@cardiff.ac.uk



Evolving Research Themes Across Two Decades (2005–2025)



How can we automatically analyze contribution?

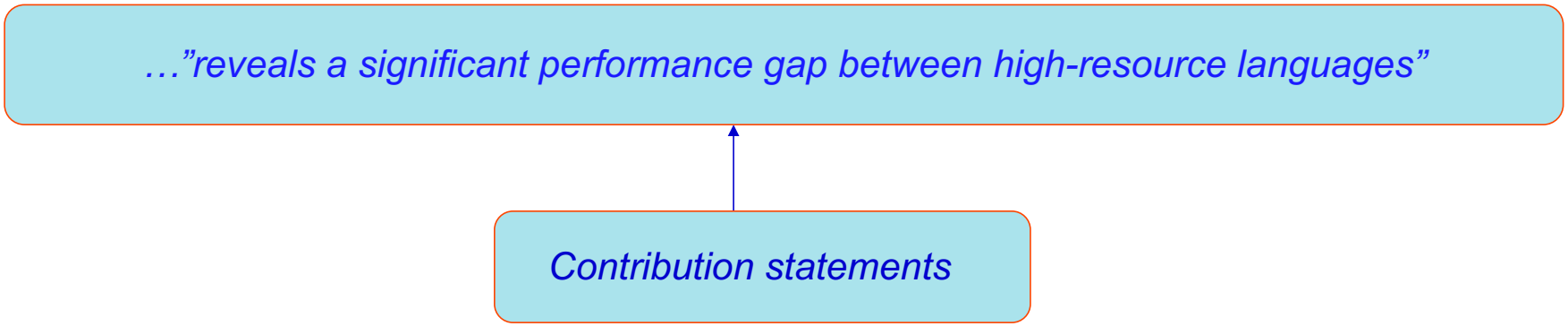
AfricaNLPContribution Dataset

A Corpus for Automatically Classifying Research Contributions in African NLP (2005–2025)

What is Contribution Statement?

Contributions *are new scientific achievements attributed to the authors*

Our evaluation reveals a significant performance gap between high-resource languages (such as English and French) and low-resource African languages." – Adelani et al. (2025)



What is Contribution Statement?

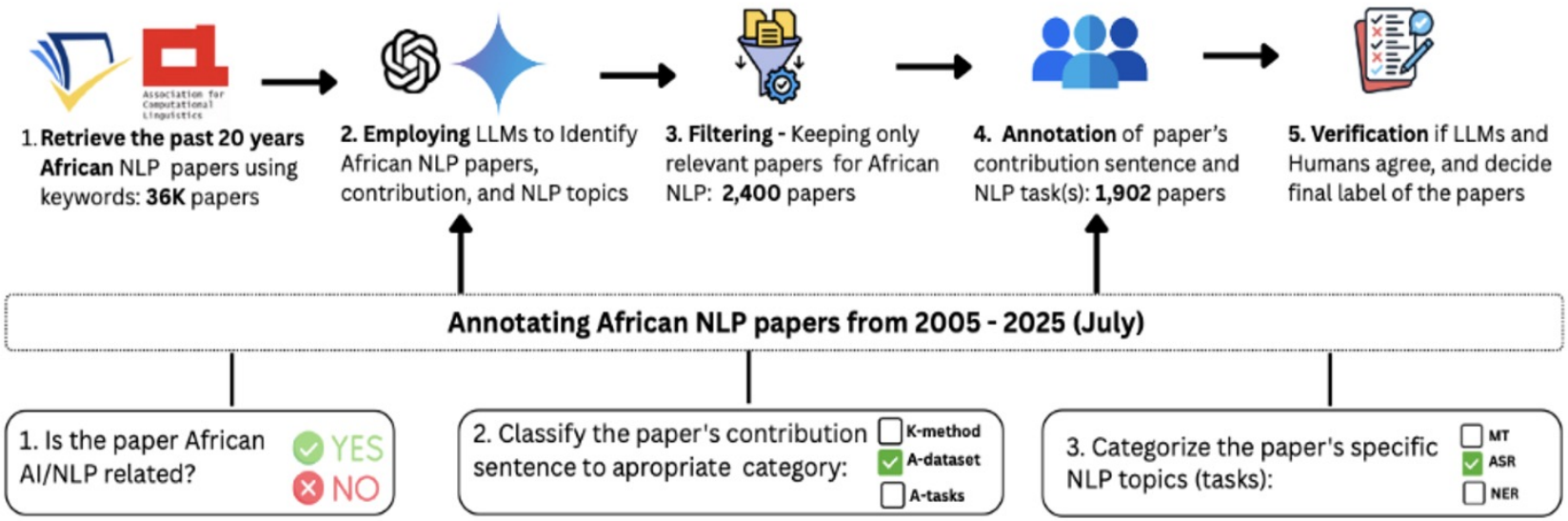
Knowledge Contribution: *Our evaluation reveals a significant performance gap between high-resource languages (such as English and French) and low-resource African languages." – Adelani et al. (2025)*

Dataset Contribution: *We release two Setswana LLM-translated benchmarks, MMLU-tsn and GSM8K-tsn, to measure Setswana knowledge and reasoning capabilities." – Brown and Marivate (2025)*

We annotated contribution statements from the abstracts of AfricaNLP research papers.

AfricaNLP Contribution Dataset

A Corpus for Automatically Classifying Research Contributions in African NLP (2005–2025)



AfricaNLPContribution Dataset

Knowledge Contributions

Advancing human understanding
(e.g., discovering the structure of DNA)

Artifact Contributions

Creating new usable systems
(e.g., ChatGPT: a general-purpose chat model)

Pramanick, A. et.al (2024). The nature of NLP: Analyzing contributions in NLP papers.

AfricaNLP Contribution Dataset

A Corpus for Automatically Classifying Research Contributions in African NLP (2005–2025)

Type	Description	Example
k-dataset	Describes new knowledge about datasets, such as their new properties or characteristics.	"Our evaluation reveals a significant performance gap between high-resource languages (such as English and French) and low-resource African languages." – Adelani et al. (2025)
k-language	Presents new knowledge about language, such as a new property or characteristic of language.	"When one homophone character is substituted by another, there will be a meaning change and it is against the Amharic writing regulation." – Belay, Tadesse Destaw and Ayele, Abinew Ali and Gelaye, Getie and Yimam, Seid Muhie and Biemann, Chris (2021)
k-method	Describes new knowledge or insights about NLP models or methods.	"This study investigates the effectiveness of Language-Adaptive Fine-Tuning (LAFT) to improve SA performance in Hausa." – Sani et al. (2025)
k-people	Presents new knowledge about people, humankind, society, or communities.	"One such group, Creole languages, have long been marginalized in academic study, though their speakers could benefit from machine translation (MT)." – Robinson et al. (2024)
k-task	Knowledge/insights about existing tasks or problem domains.	"These results emphasize the importance of selecting appropriate pre-trained models based on linguistic considerations and task requirements." – Aali et al. (2024)
a-dataset	Introduces a new NLP dataset (i.e., textual resources such as corpora or lexicon).	"Last, we release two Setswana LLM-translated benchmarks, MMLU-tsn and GSM8K-tsn, to measure Setswana knowledge and reasoning capabilities." – Brown and Marivate (2025)
a-method	Introduces or proposes a new or novel NLP methodological approach or model to solve NLP task(s).	"We design and implement a new MMT model framework suitable for our new generated dataset." – Xiao et al. (2025)
a-task	Introduces or proposes a new or novel NLP task formulation (i.e., well-defined NLP problem).	"In this work, we introduce the challenge of Meroitic decipherment as a computational task" – Otten and Anastasopoulos (2025)

Table 3: Description of the taxonomy for NLP research Knowledge (**k**) and artifacts (**a**) contributions with examples from the `AfricaNLPContributions` dataset.

The Rise of African NLP: Contributions, Contributors, and Community Impact (2005–2025) Belay, Tadesse et al. 2025.

Why AfricaNLPContribution Dataset?

A Large-Scale Corpus for Automatically Classifying Research Contributions in African NLP (2005–2025)

Automatically detecting contributions in research papers allows us to:

Understand what progress has already been made

Reveal gaps and opportunities for future research

Benchmarking Contribution Classification

Automatically Classifying NLP Research Contributions

Goal: Automatically classify each contribution sentence into one of 8 predefined classes (5 knowledge + 3 artifact classes)

SciBERT and **AfroXLM-R** perform best across most contribution types.

Contrib.	BERT	SciBERT	AfroXLMR	GPT*
k-language	**0.69**	0.68	0.65	0.49
k-method	0.79	**0.80**	0.78	0.63
k-people	0.42	0.49	**0.54**	0.49
k-task	**0.41**	0.40	0.36	0.30
k-dataset	0.54	0.51	**0.62**	0.55
a-dataset	0.80	0.81	**0.83**	0.80
a-method	0.84	**0.85**	0.84	0.76
a-task	0.42	**0.48**	0.43	0.47
Overall	0.61	**0.63**	**0.63**	0.50

Benchmarking Contribution Classification

Automatically Classifying NLP Research Contributions

SciBERT and **AfroXLM-R** perform best across most contribution types.

SciBERT performs better in *k-method* and *a-method classes*, benefiting from scientific-domain pretraining.

Contrib.	BERT	SciBERT	AfroXLMR	GPT*
k-language	**0.69**	0.68	0.65	0.49
k-method	0.79	**0.80**	0.78	0.63
k-people	0.42	0.49	**0.54**	0.49
k-task	**0.41**	0.40	0.36	0.30
k-dataset	0.54	0.51	**0.62**	0.55
a-dataset	0.80	0.81	**0.83**	0.80
a-method	0.84	**0.85**	0.84	0.76
a-task	0.42	**0.48**	0.43	0.47
Overall	0.61	**0.63**	**0.63**	0.50

Benchmarking Contribution Classification

Automatically Classifying NLP Research Contributions

SciBERT and **AfroXLM-R** perform best across most contribution types.

SciBERT performs better in *k-method* and *a-method classes*, benefiting from scientific-domain pretraining.

AfroXLM-R performs strongly on k-dataset and a-dataset, reflecting better regional and linguistic adaptation.

Contrib.	BERT	SciBERT	AfroXLMR	GPT*
k-language	**0.69**	0.68	0.65	0.49
k-method	0.79	**0.80**	0.78	0.63
k-people	0.42	0.49	**0.54**	0.49
k-task	**0.41**	0.40	0.36	0.30
k-dataset	0.54	0.51	**0.62**	0.55
a-dataset	0.80	0.81	**0.83**	0.80
a-method	0.84	**0.85**	0.84	0.76
a-task	0.42	**0.48**	0.43	0.47
Overall	0.61	**0.63**	**0.63**	0.50

The Paradox of Progress: AfricaNLP in the LLM Era

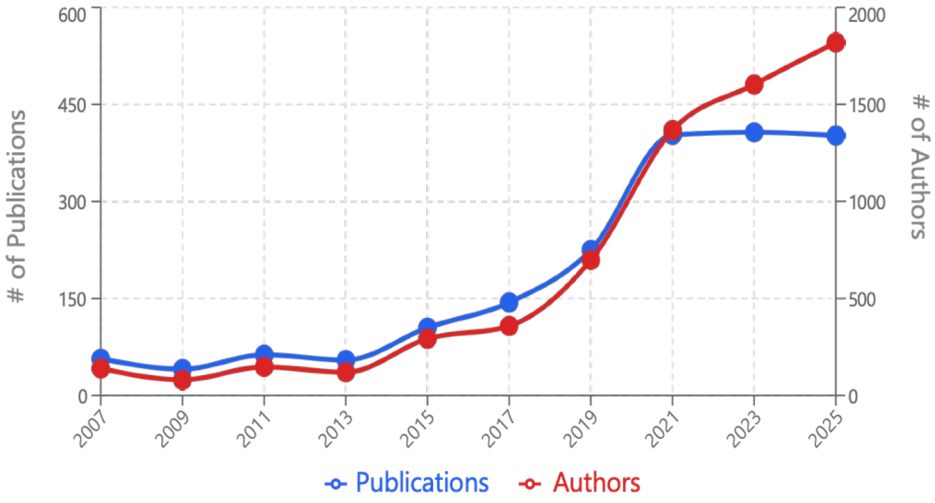
Fast growth, slow inclusion

Two decades of steady growth in AfricaNLP research

Yet, foundation models still under-represent African languages and cultures

Only few languages are supported across models.

Progress at the periphery — **exclusion at the core** of global AI.



The State of Large Language Models for African Languages: Progress and Challenges. (Hussen, K.Y., et.al 2025) – ***Best Paper Deep Learninng Indaba 2025***

Outline

A Two-Decade Journey of African NLP

The Illusion of Inclusion

THE TRIPLE GAP

Pretraining Data: Biased and Noisy,
Evaluation Data: Inaccurate and Unreliable,
Culture Missing: The Cultural Blind Spot of LLMs

Pretraining Data: Biased and Noisy

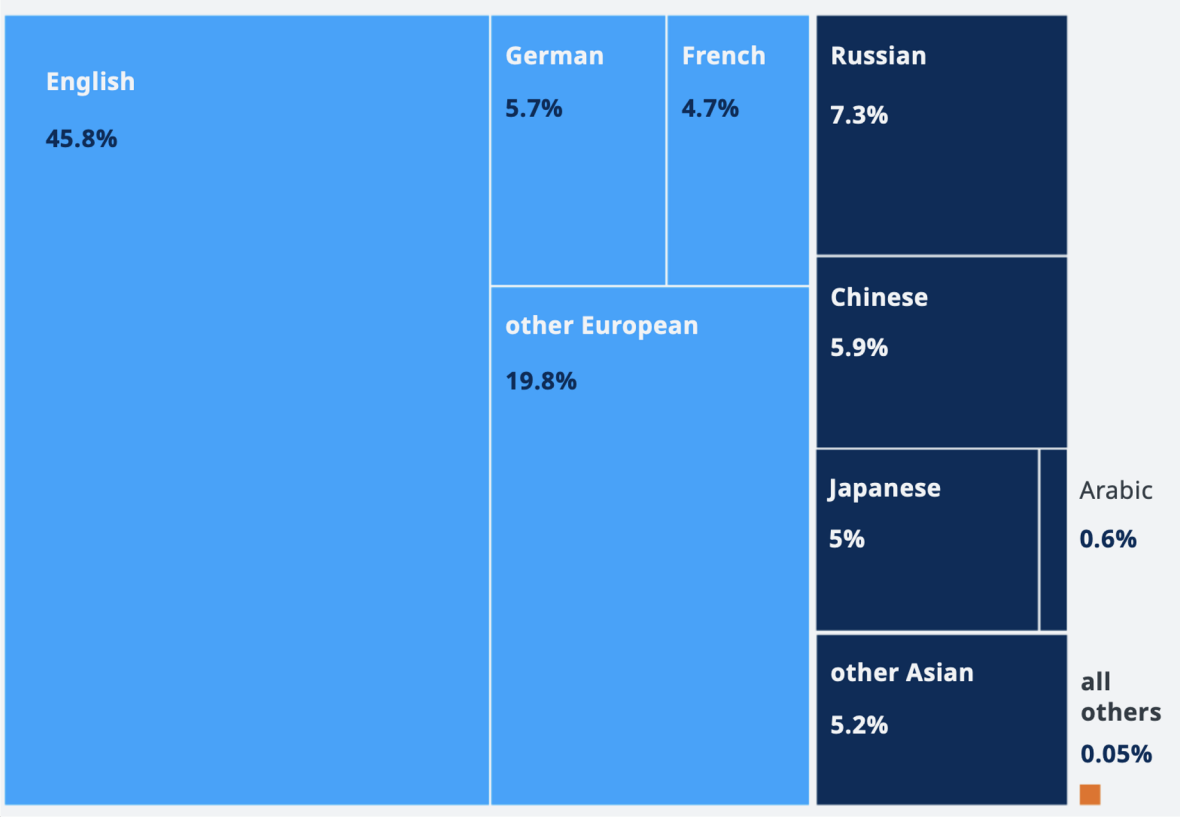
Pretraining Data — Biased and Noisy

Scraped from the web, but not from the world

- Most multilingual corpora (e.g., CCAligned, OSCAR, mC4) contain **<1 % African language data**.
- Sources are noisy**,** domain-limited, and often translated or duplicated**.**
- African languages become **statistically invisible** during LLM pretraining.

Pre-training Data — Biased and Noisy

languages in the Common Crawl internet archive



Source: Common Crawl | More Info: github.com/dw-data/ai-languages

30%

World languages are African (Ethnologue)

0.05%

Pretraining Data — Biased and Noisy

Scraped from the web, but not from the world

- We manually audit the quality of 205 language-specific corpora
  - (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4).
- African languages corpora have systematic issues:
  - At least 15 corpora have no usable text, and a significant fraction contains less than 50% sentences of acceptable quality
  - Many are mislabeled

….Representation Washing

Kreutzer et al. (2022). *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets.* TACL, 10: 50–72.

Pretraining Data — Biased and Noisy

Representation Washing

- Community may feel a sense of progress and growing equity, despite the actual quality of the resources for these languages.
- If low-quality datasets are used as benchmarks they may exaggerate model performance, making low-resource NLP appear more solved than it is

These effects could result in productive effort being redirected away from these languages.

Kreutzer et al. (2022). *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets.* TACL

Pretraining Data — Biased and Noisy

Script and Tokenization Bias

- Africa has 23 writing systems: LLM tokenizers support only 3 (Latin, Arabic, Geʻez) appear in model tokenizers.
- Most African scripts are invisible to LLMs
- Missing scripts → broken tokenization, loss of meaning, and semantic noise during training

The State of Large Language Models for African Languages: Progress and Challenges. (Hussen, K.Y., et.al 2025)

Evaluation Data Inaccurate and Unreliable

Evaluation Data — Inaccurate and Unreliable

"Gold Standards" Misjudge African Language Performance

- Widely used multilingual benchmarks often or mis-represent African text.
- Evaluation sets are frequently machine-translated
- As a result, benchmark-driven claims often misrepresent true model capability on African languages.

Abdulmumin et al. (2024). *Correcting FLORES Evaluation Dataset for Four African Languages.* WMT 2024.

Evaluation Data — Inaccurate and Unreliable

"Gold Standards" Misjudge African Language Performance

- FLORES is widely used to evaluate MT for African languages.
- But in four languages we audited — Hausa, Xitsonga, Sesotho, and isiZulu, we found translation errors, inconsistencies, and of machine-generated text.
- This means model performance has been misreported for years.

If the evaluation set is broken, then claims of model performance are also broken.

Correcting FLORES evaluation dataset for four African languages — Abdulmumin et al. (2024)

Evaluation Data — Inaccurate and Unreliable

"Gold Standards" Misjudge African Language Performance

…future translation efforts, particularly in low-resource languages, prioritize the active involvement of native speakers at every stage of the process to ensure linguistic accuracy and cultural relevance.

Abdulmumin et al. (2024). *Correcting FLORES Evaluation Dataset for Four African Languages.* WMT 2024.

Evaluation Data — Inaccurate and Unreliable

The Missing Benchmarks

- Most tasks focus on simple classification: very few evaluate generation, QA, reasoning, or dialog

- Lack of shared benchmarks means progress is often non-comparable, inconsistent, or anecdotal

Missing Culture — The Cultural Blind Spot of LLMs

Fluent in Language, Blind in Meaning

- LLMs can produce African text but lack cultural understanding
- Linguistic fluency ≠ cultural understanding.
- Cultural errors lead to harmful stereotypes

Myung et al. (2024). *BLEND: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages.* NeurIPS 2024.

Missing Culture — The Cultural Blind Spot of LLMs

Fluent in Language, Blind in Meaning

- **BLEND**
  - Comprises 52.6k question-answer pairs from 16 countries/regions
  - GPT-4 ≈ 79 % accuracy on U.S. culture but only 12 % on Amharic/Hausa.
  - LLMs perform better in English than the local languages.

Myung et al. (2024). *BLEND: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages.* NeurIPS 2024.

THE TRIPLE GAP

Pretraining Data: Biased and Noisy,

Evaluation Data: Inaccurate and Unreliable,

Culture Missing: The Cultural Blind Spot of LLMs

Addressing these requires clean and diverse pretraining data, reliable evaluation benchmarks, and culturally grounded understanding.

Bridging the Triple Gap: Data, Evaluation, Culture

Dimension	What's Missing	Consequence
Pretraining	Clean, diverse data	Bias propagation
Evaluation	Reliable benchmarks	Distorted performance
Culture	Cultural knowledge & context	Culturally incoherent outputs

Outline

A Two-Decade Journey of African NLP

The Illusion of Inclusion

Selected Works

Selected Research in African Languages

Computational Social Science, Low-resource, Multilingual NLP

- BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages (Muhammad et al.,2025) ACL2025 Best Resource paper Award
- AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages (Muhammad et al., 2025)
- AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text (Belay et al., 2025)
- Word Translation in the Era of Large Language Models (Muhammad, et.al 2025)

BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages

Shamsuddeen Hassan Muhammad*, Nedjma Ousidhoum*, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, Saif M. Mohammad

*: Equal contribution. **Contact**: s.muhammad@imperial.ac.uk, OusidhoumN@cardiff.ac.uk

This project has received two awards

ACL 2025 Best Resource Paper | SemEval 2025 Best Task Award

🤗: BRIGHTER Emotion Categories Dataset | 🤗 HF: BRIGHTER Emotion Intensities Dataset

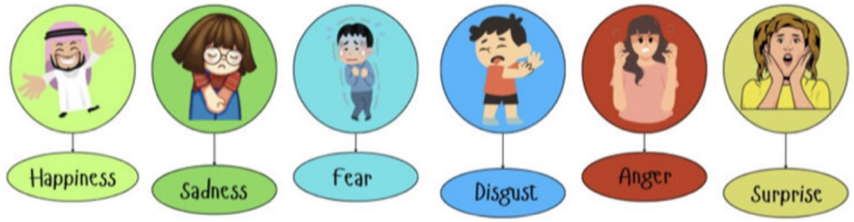
BRIGHTER Dataset Paper | SemEval-2025 Task 11 Paper | BRIGHTER Experimental Code

SemEval2025-Task-11 Repo | BRIGHTER Paper Slides | SemEval2025-Task-11 Slides

https://brighter-dataset.github.io/

Motivation

- **Human communication**
  is deeply emotional



- **Multilingual and cultural challenges**
  Emotional expression varies across **languages, cultures,** and **contexts** (perceived subjectively).

Focuses on **perceived emotions**

Predict… *what emotion most people will think the speaker may be feeling, given a sentence or a short text snippet uttered by the speaker.*

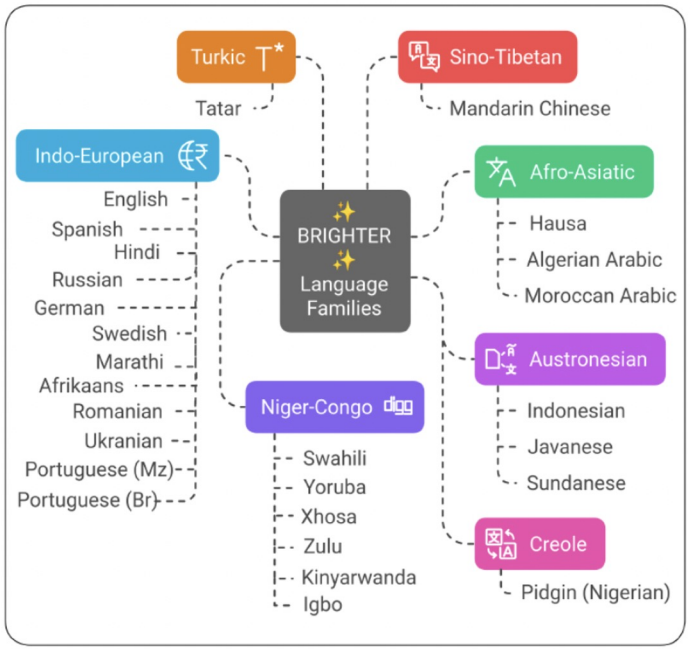
BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages

Task: perceived emotions, i.e., what emotion most people think the speaker might have felt given a sentence or a short text snippet uttered by the speaker.

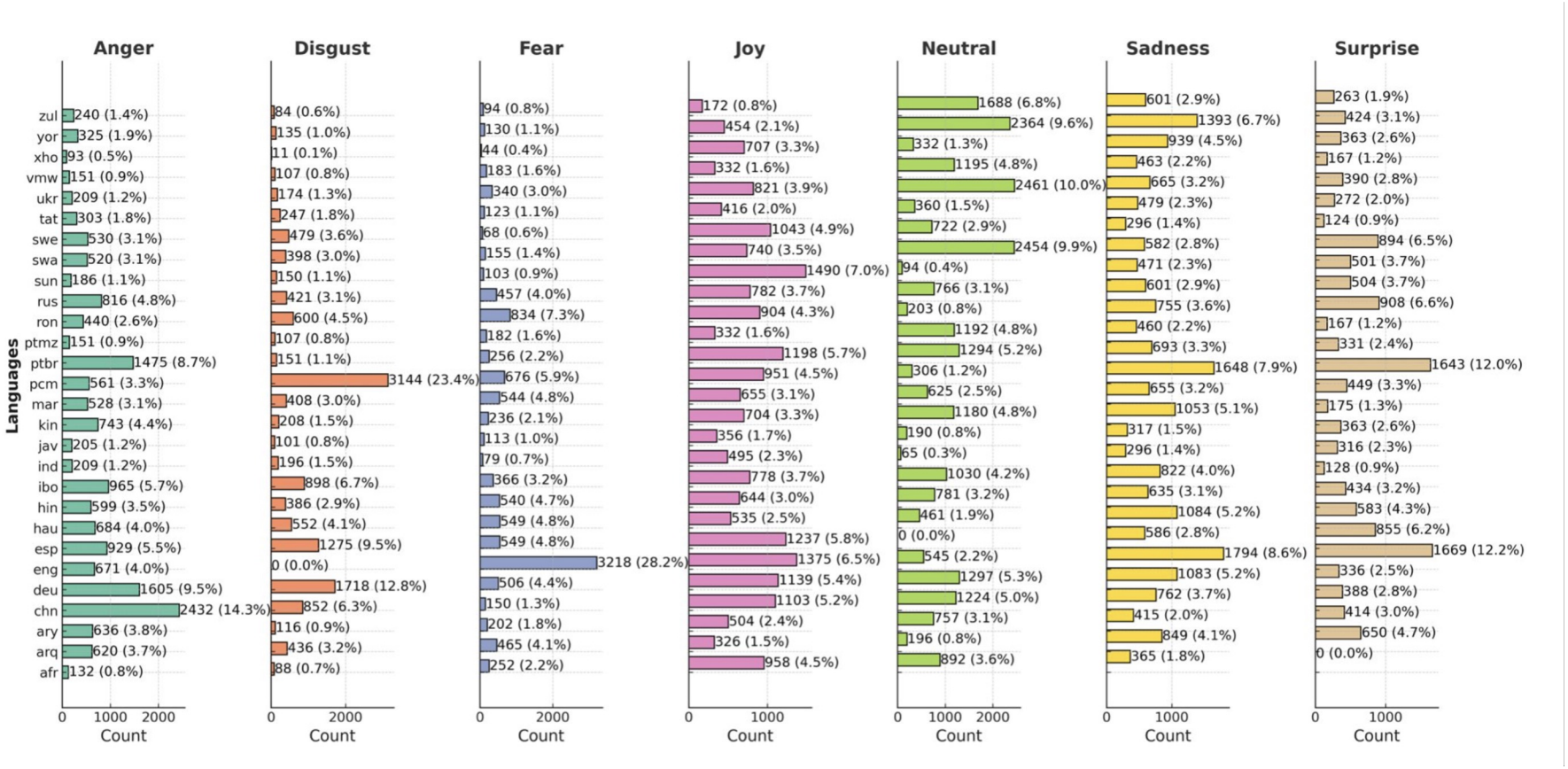
Approximately 100,000 instances from diverse data sources: speeches, social media, news, literature, and reviews.

From Africa, Asia, Eastern Europe, Latin America (28 languages)

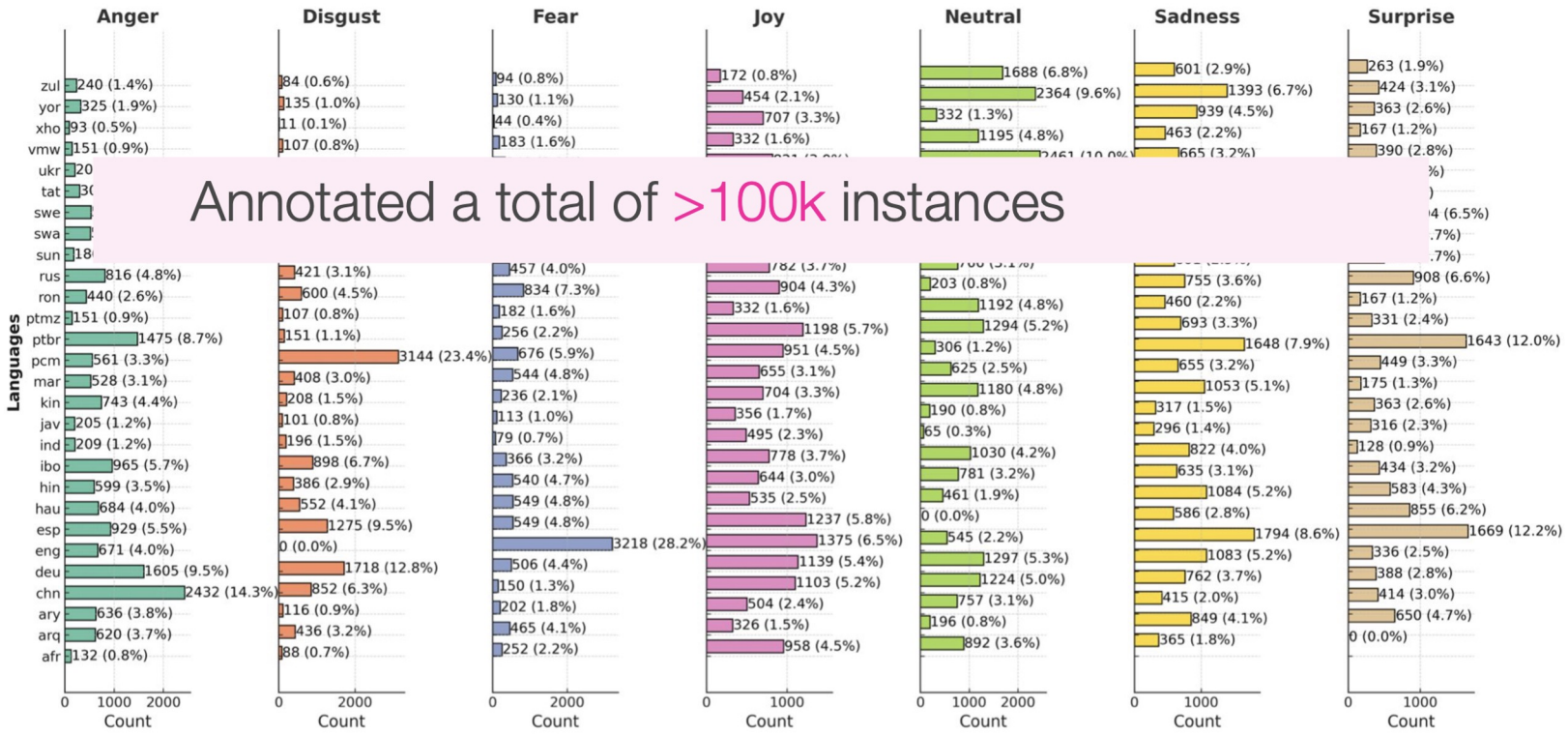
***Predominantly** from Africa 12 out of 28*



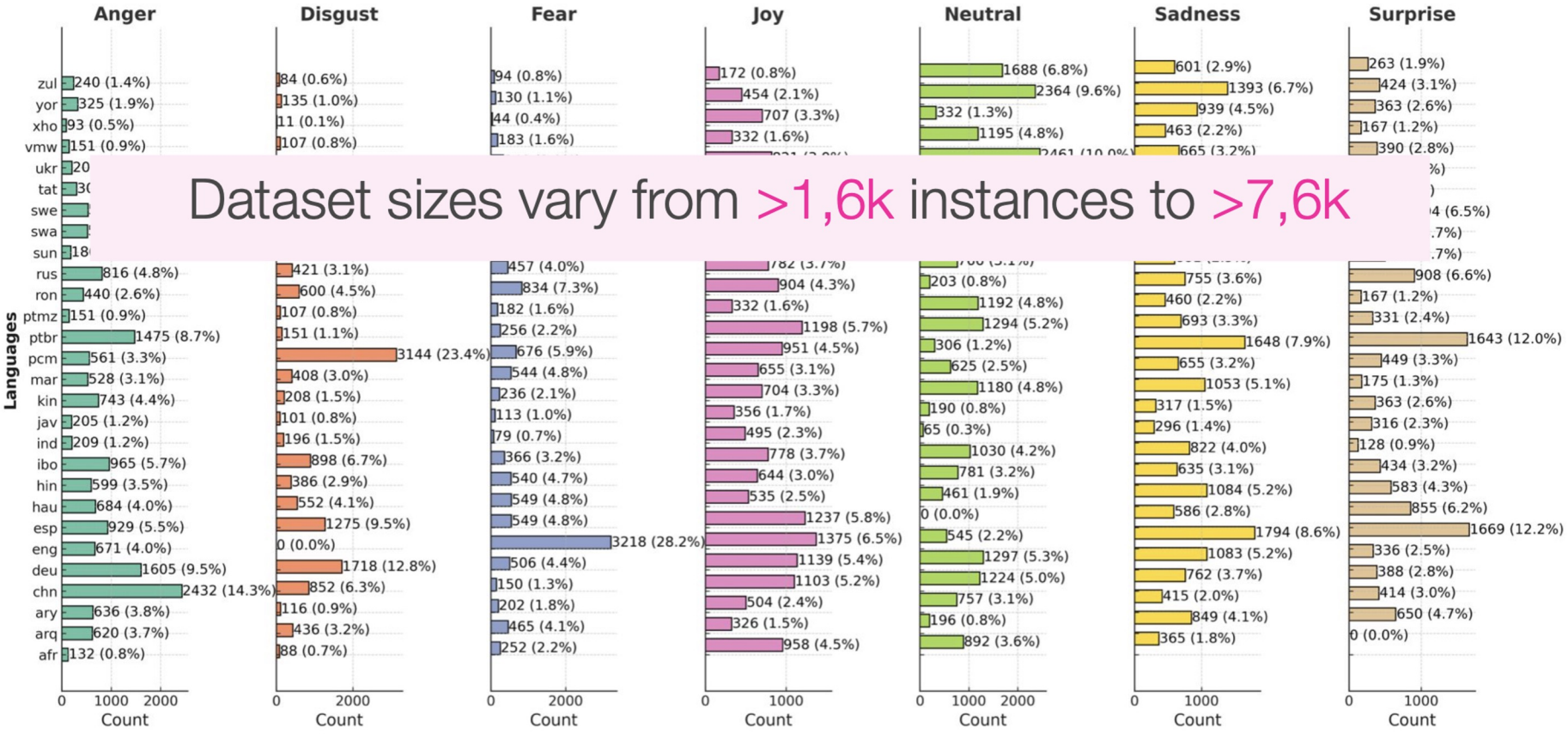
BRIGHTER Final Dataset



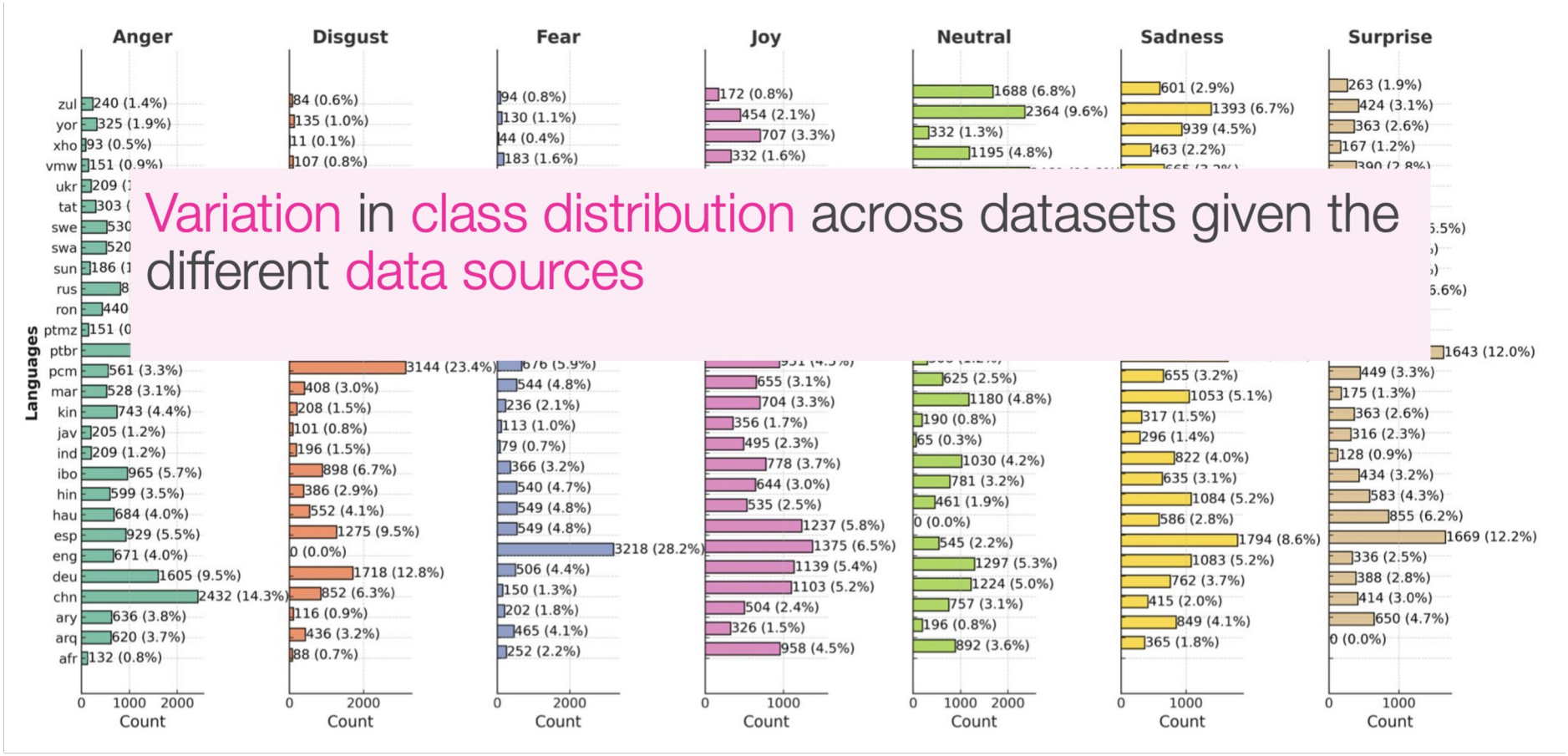
BRIGHTER Final Dataset



BRIGHTER Final Dataset



BRIGHTER Final Dataset

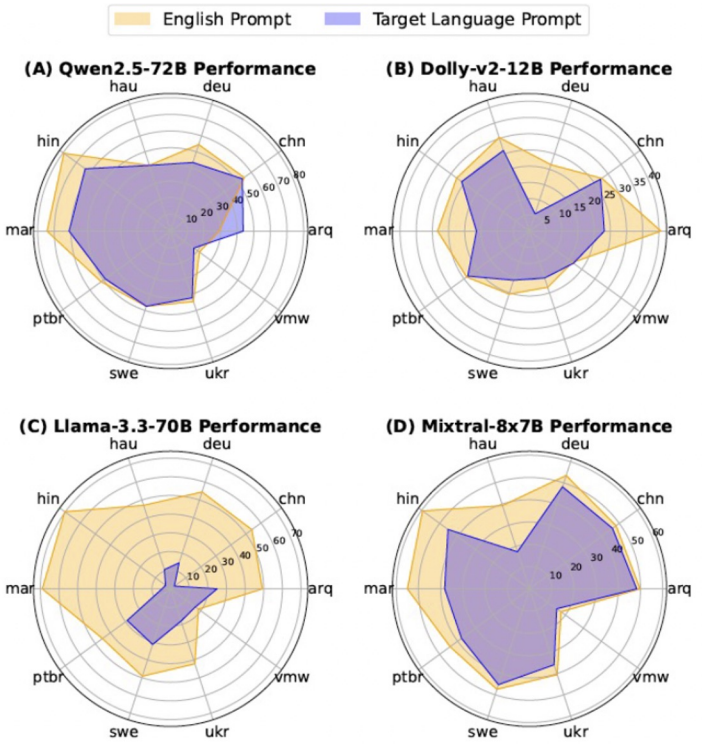


Prompting in English vs. Target Language

We prompt LLM in both English and Target Language

Overall, models tend to perform better when **prompted in English**

Except **for Algerian Arabic**, for which **Qwen2.5-72B** performs better



BRIGHTER

***LLMs** still struggle with predicting perceived emotions and their intensity levels*

***LLM performance** is highly dependent on the **wording** of the prompt, its **language**, and the **number of shots** in few-shot settings.*

Word Translation in the Era of Large Language Model

- LLM show good performance for MT at sentence and document-level
- Word translations are crucial for low-resource language NLP
- Evaluated LLMs (closed vs. open-source)
- 106 languages, including many low-resource ones
- Benchmark against **PanLex** + Google Translate

Word Translation in the Era of Large Language Model, Muhammad, et.al. 2025

Word Translation in the Era of Large Language Model

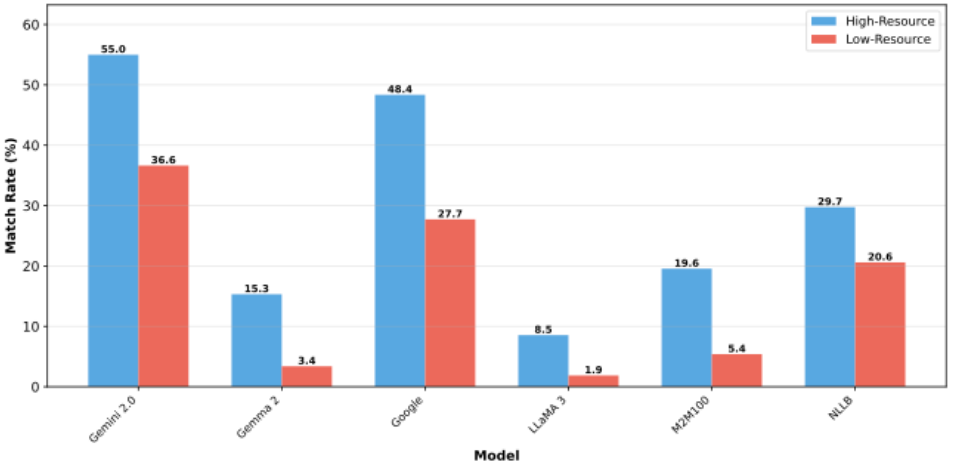


Figure 4: Performance comparison between high-resource and low-resource languages (Joshi et al., 2020)

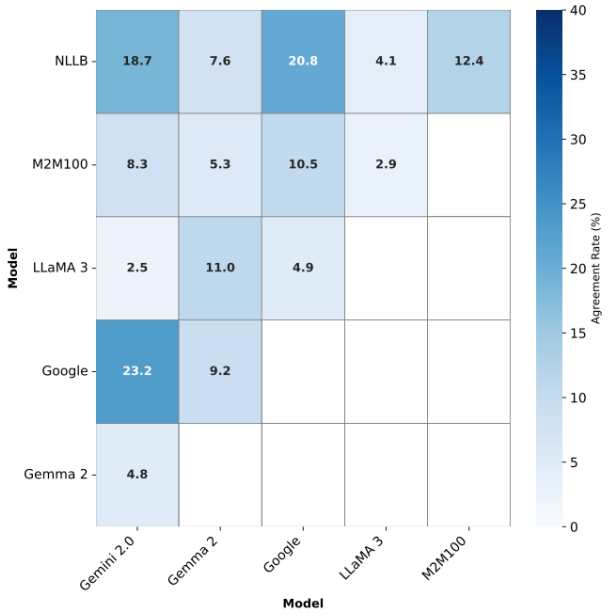


Figure 2: Pairwise agreement analysis reveals generally low alignment between LLMs and traditional MT systems, with the notable exception of the high agreement between the closed-source Gemini and Google Translate.

Beyond General LLMs

Domain-Specific Continual Pre-Training

Beyond General LLMs

Smaller, fine-tuned models outperform large LLMs on African tasks

General LLMs underperform on African languages due to limited exposure and poor cultural alignment.

Low-resource status leads to high untranslated, invalid, or hallucinated outputs.

Continual pretraining on African-specific corpora (e.g., AfriSocial) significantly improves accuracy and reliability.

Gururangan et al., ACL 2020 — Don't Stop Pretraining: Adapt LMs to Domains and Tasks.

AfriSocial Corpus

A Large-Scale Social Media Corpus for African Languages

Corpus Composition

- Covers 19 African Languages
- Social Media + News Articles
- Over millions of tokens, spanning public discourse, informal text, and formal reporting

Belay et al. (2025). *AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text.*

AfroXLMR-Social: Continual Pretraining for African NLP

Smaller, Smarter, and More Culturally Grounded Than Giant LLMs

- AfroXLMR-Social
  - Trained on **AfriSocial Corpus** (news + social media in 19 languages)
- Uses:

  Domain-Adaptive Pretraining (DAPT): tunes model on African data

  Task-Adaptive Pretraining (TAPT): tailors to specific downstream tasks

Belay et al. (2025). *AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text.*

AfroXLMR-Social

Smaller, Smarter, and More Culturally Grounded Than Giant LLMs

AfroXLMR-Social:

AfroXLMR-Social consistently outperforms ***GPT-4o***, ***Gemini***, *and* ***Llama-3*** on African-language tasks.

Continual pretraining remains vital for low-resource, culturally rich languages.

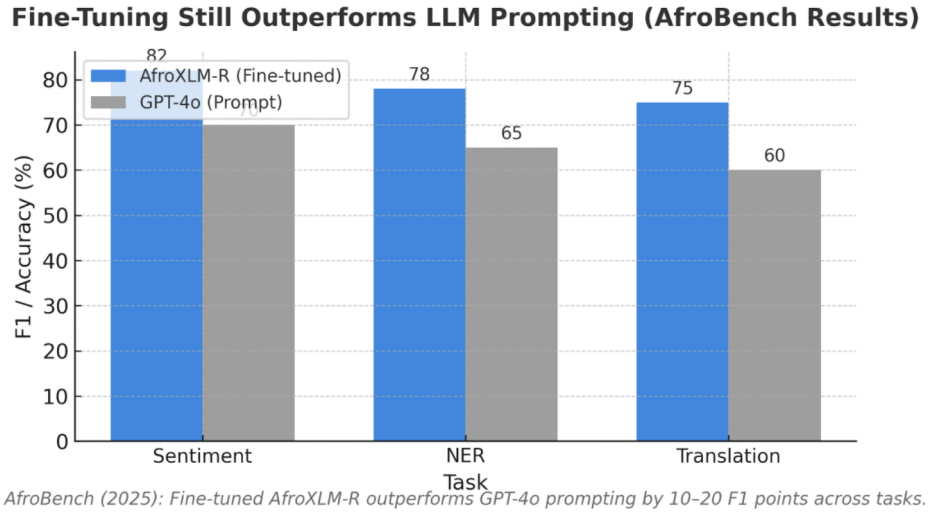
Belay et al. (2025). *AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text.*

AfroBench

Smaller, fine-tuned models outperform large LLMs on African tasks

AfroBench : Fine-tuned **AfroXLM-R**, **AfriTeVa**, and **NLLB** outperform LLMs like GPT-4o, Gemini, and Llama-3 on multiple African-language tasks.

Bigger isn't always better — AfroBench shows that contextual fine-tuning still beats prompting.



AfroBench (2025): Fine-tuned AfroXLM-R outperforms GPT-4o prompting by 10–20 F1 points across tasks.

Ojo et al. (2023). *AfroBench: How Good are Large Language Models on African Languages?* ACL 2023.

Scaling Isn't Enough — Roadmap to Real Progress

Progress for African LLMs must begin with linguistic infrastructure—not parameter inflation.
Before we scale up models, we must scale up languages.

Collaboration Drives Lasting Progress

Grassroots communities are the true infrastructure of African NLP

We need more community-led efforts to ensure languages are not left behind

Inclusion is not automatic. We must design for it — together.

Thank You

https://shmuhammadd.github.io/

s.Muhammad@imperial.ac.uk