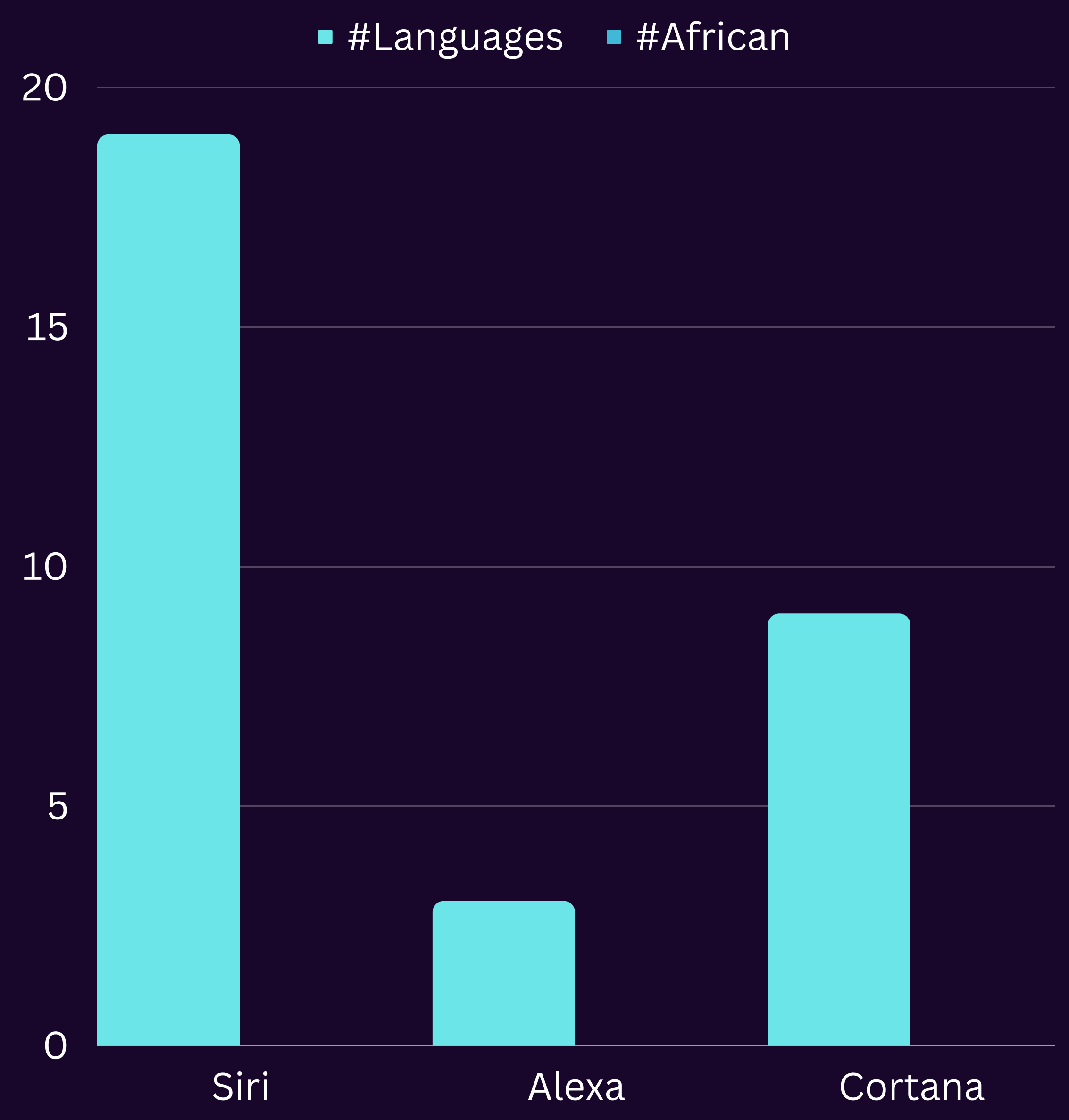
BONAVENTURE F. P. DOSSOU

BRIDGING LINGUISTIC FRONTIERS: MACHINE LEARNING & NLP INNOVATIONS EMPOWERING AFRICAN LANGUAGES: CHALLENGES, PROGRESS, AND PROMISING FUTURES

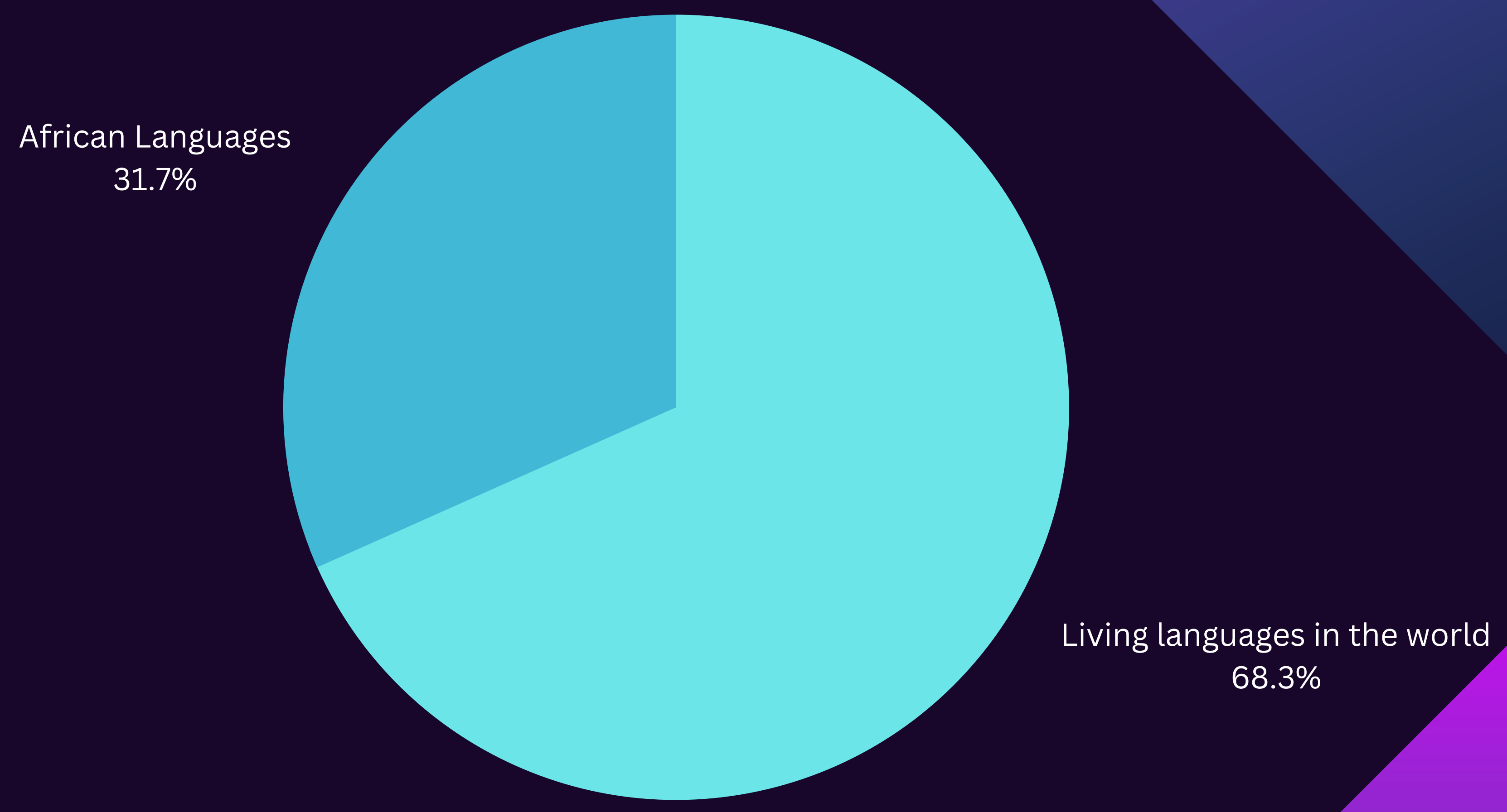
https://bonaventuredossou.github.io/

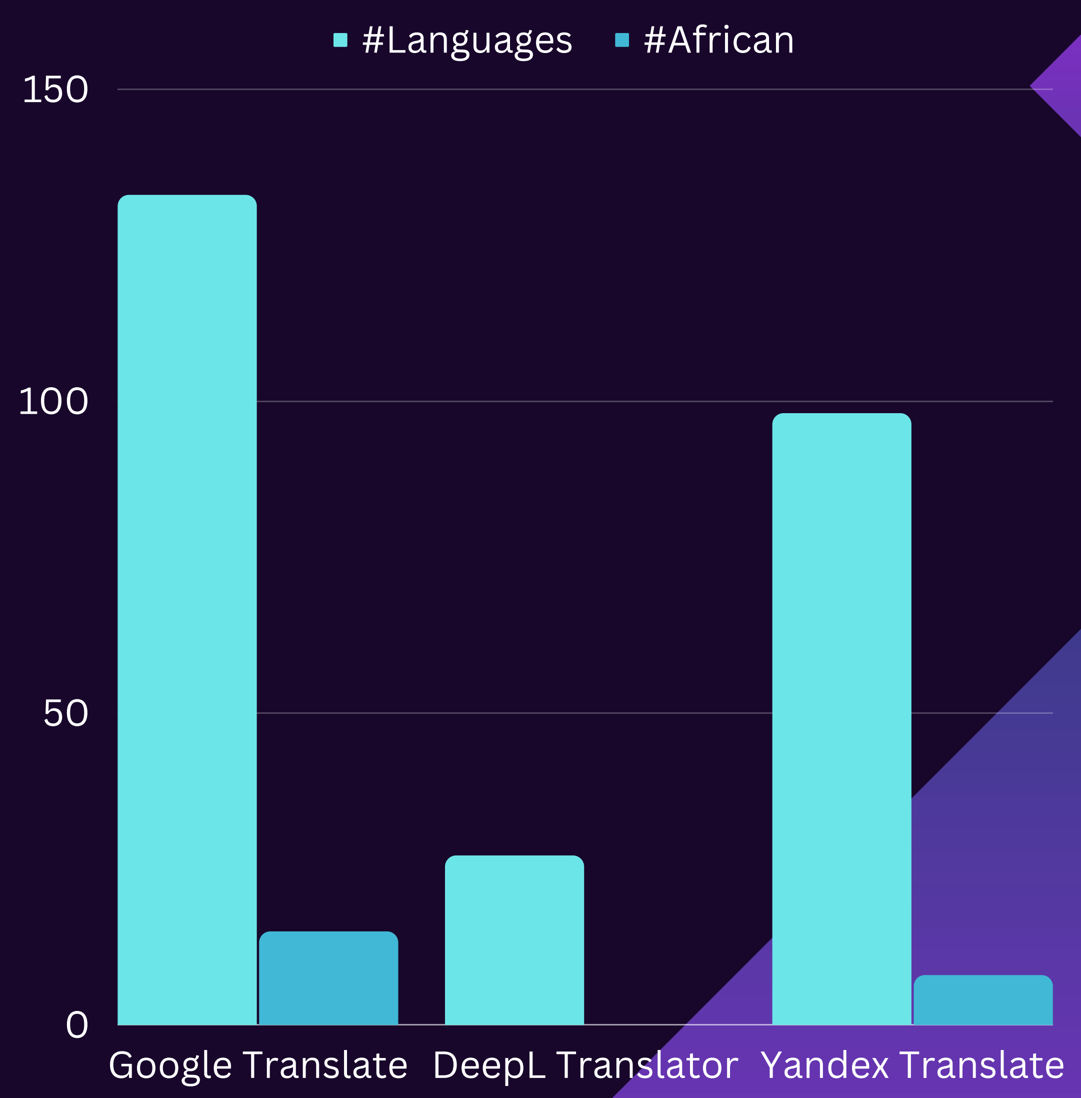
Ph.D. Student, McGill University & Mila Quebec AI Institute
Research Scientist, Lelapa AI

CHALLENGE: LACK OF INTEGRATION

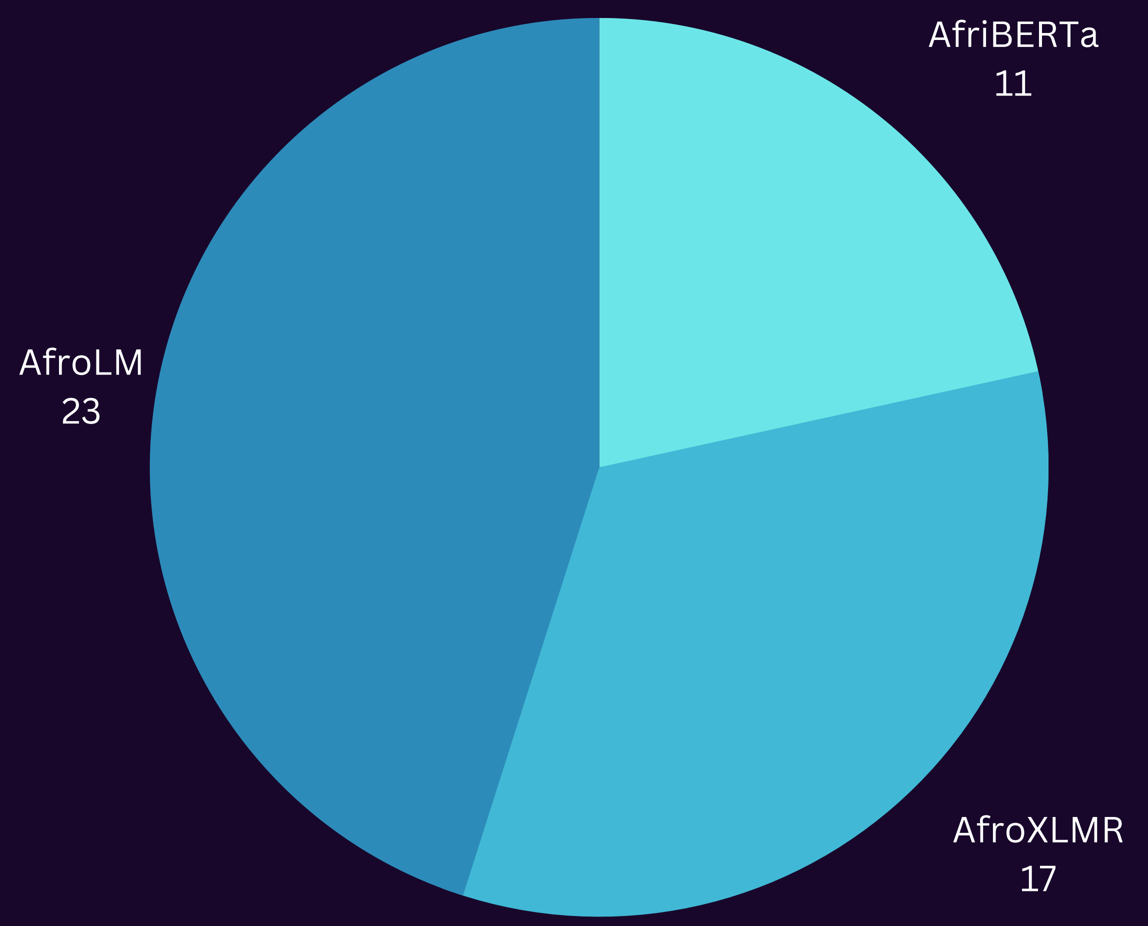


# of Supported African languages in existing Voice Assistants

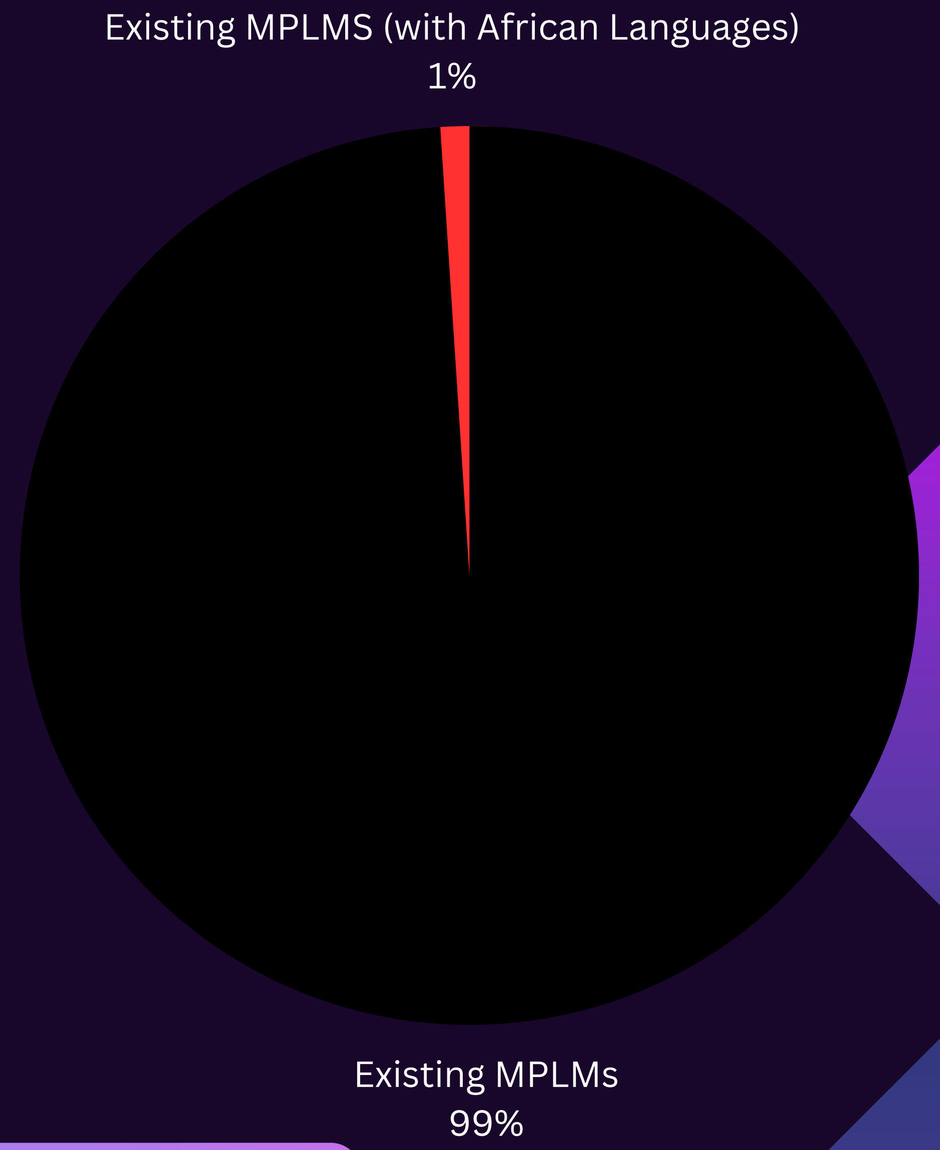




CHALLENGE: LACK OF REPRESENTATION & DATA

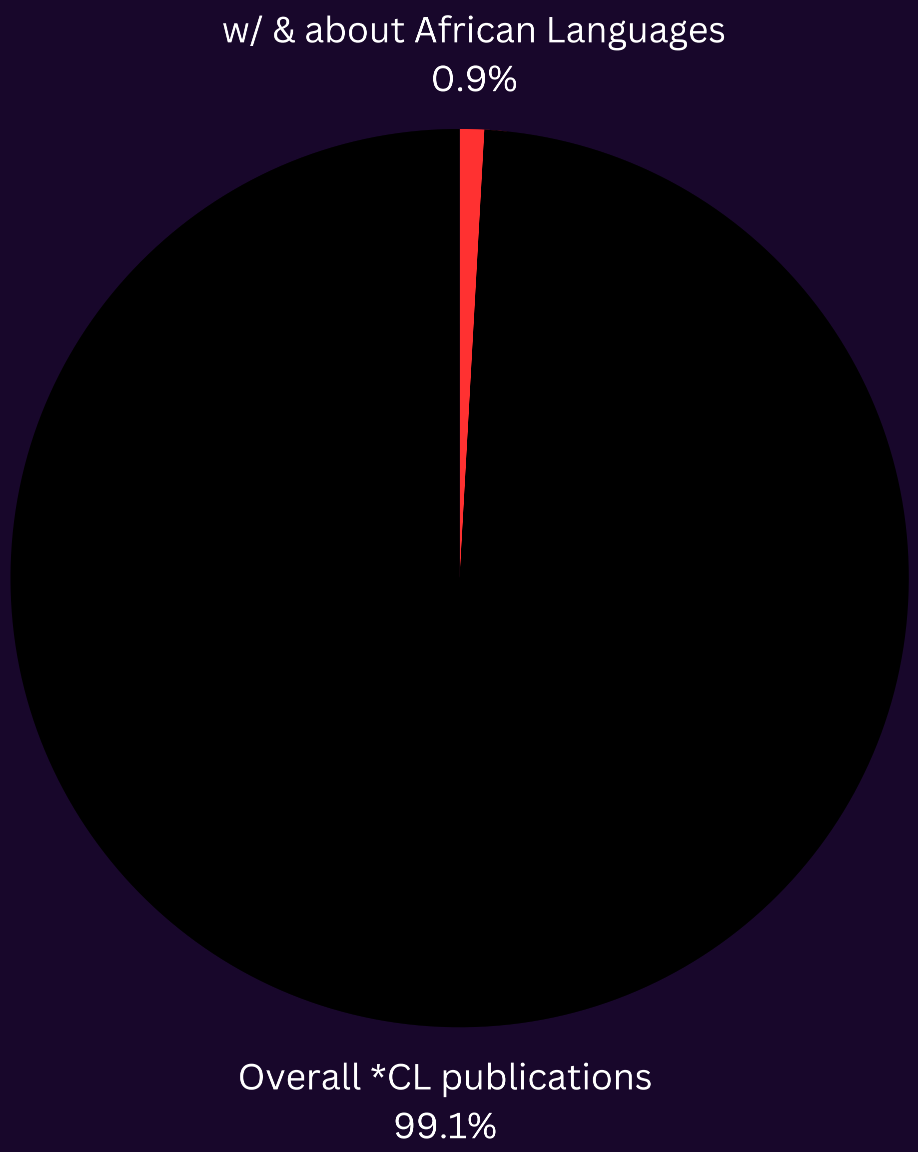


Existing Entirely Africa-centric MPLMs



Existing language systems including MPLMs (data from HuggingFace)

CHALLENGE: LACK OF REPRESENTATION & DATA



Existing Datasets (data from HuggingFace)

*CL Publications

PROGRESS: BUILDING MORE AFROCENTRIC DATASETS

- 3 versions of FFR (Fon-FRench) Dataset for Fon-French machine translation (ICLR & ACL 2020)
- 2 versions of MasakhaNER (NER datasets of 10 & 20 languages respectively - TACL 2021/EMNLP 2022)
- MasakhaNEWS (News dataset of 16 languages - ICLR 2023)
- MasakhaPOS (POS dataset of 20 languages - ACL 2023)
- NaijaSenti (Sentiment Analyses of 4 Nigerian languages - LREC 2022)
- YOSM (Yorùbá Sentiment Corpus for Nollywood Movie Reviews - ICLR 2022)
- BibleTTS (TTS dataset for 10 African Languages - Interspeech 2022)
- AfriSpeech (200hr Pan-African speech corpus for clinical and general for 120 African accents - TACL 2023)
- AfriQA (Cross-lingual QA dataset for 9 African languages - EMNLP 2023)

PROGRESS: LANFRICA - INCREASING DISCOVERABILITY OF AFRICAN LANGUAGES RESOURCES

LANFRICA

CONNECTING ALL AFRICAN LANGUAGE RESOURCES

A home of African Languages Resources

PROGRESS: LANFRICA - INCREASING DISCOVERABILITY OF AFRICAN LANGUAGES RESOURCES

LANFRICA

CONNECTING ALL AFRICAN LANGUAGE RESOURCES

A home of African Languages Resources

www.lanfrica.com

01 Unique marketplace for the creator-community: we license the commercial use of the data, and we give lifelong benefits to the data creators.

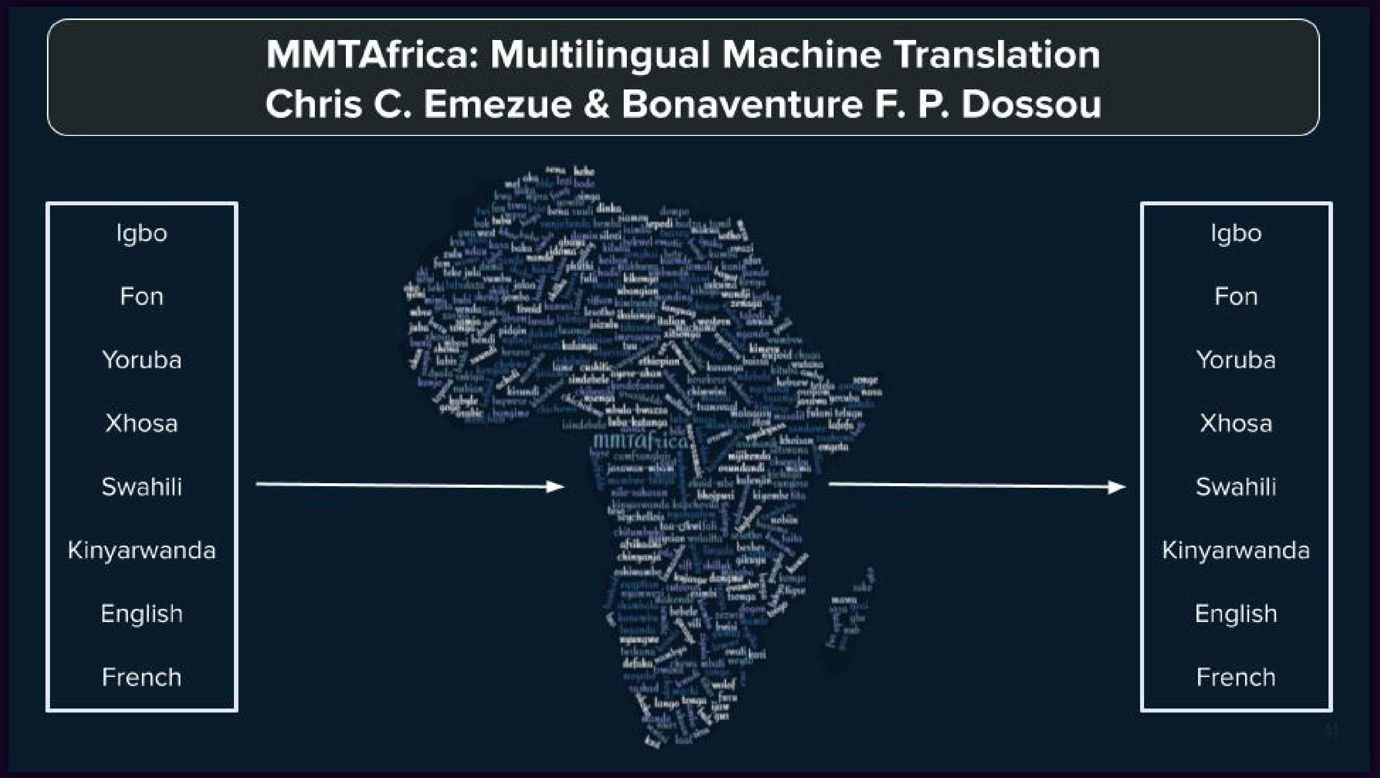
02 Annotation tools with user-base: Lanfrica has a large user base of data creators/annotators, contributing to the dataset creation, to the benefit of the community.

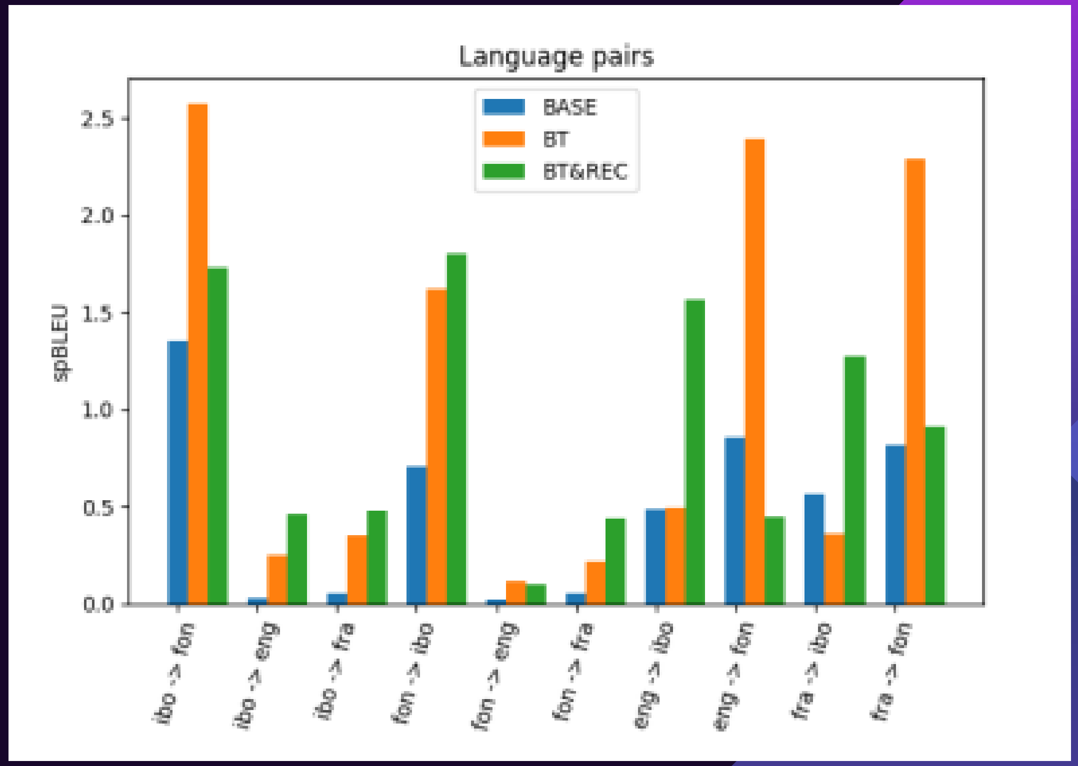
03 Central hub to find African datasets, by linking all African datasets on the web to make them accessible for free.

04 High-quality, useful African datasets to build more language technologies for African languages.

05 Make African resources more discoverable by providing a central hub to easily search/find African resources.

- MMT AFRICA
  - BASE: finetuning on the many-to-many translation task
  - BT: finetuning with back-translation
  - BT&REC: finetuning w/ joint backtranslation and reconstruction





improvements from MMTAfrica over the FLORES 101 benchmarks
(spBLEU gains ranging from +0.58 in Swahili to French to +19.46 in French to Xhosa)

PROGRESS: LEVERAGING TRANSFER LEARNING ON AFRICAN LANGUAGE NLP DOWNSTREAM TASKS

- MasakhaNER: Named Entity Recognition for African Languages
- Africa-centric Transfer Learning for Named Entity Recognition
  - Useful features that play key roles in performance improvements: transfer language dataset size, target language dataset size, geographic distance, phonological distance
- MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African languages (ACL 2023)
- MasakhaNEWS: News Topic Classification for African languages (AACL 2023)
- FonMTL: Towards Multitask Learning for the Fon Language (EMNLP 2023)
- and many more ....

PROGRESS: HOW MUCH DATA WE MINIMALLY NEED TO PERFORM WELL IN TRANSFER LEARNING SETTING

A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation

- MAFAND-MT: Masakhane Anglo & Franco Africa News Dataset for Machine Translation
- Transfer Learning Across Languages: Continual Pretraining & Many2Many Translation
- Transfer Learning Across Domains
  - REL+NEWS: Fine-tuning the aggregation of religious and news domain data
  - REL→NEWS: Training on the religious domain then finetuning on the news domain
  - REL+NEWS→NEWS: REL+NEWS, followed by additional fine-tuning on the news domain

PROGRESS: HOW MUCH DATA WE MINIMALLY NEED TO PERFORM WELL IN TRANSFER LEARNING SETTING

A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation

Zero-shot vs after finetuning Machine Translation evaluation: Finetuning HELPS !!!

	fr-xx						*en-xx*											
Model	**bam**	**bbj**	**ewe**	**fon**	**mos**	**wol**	**hau**	**ibo**	**lug**	**luo**	**pcm**	**swa**	**tsn**	**twi**	**yor**	**zul**	**AVG**	**MED**
BLEU																		
M2M-100 0-shot	–	–	–	–	–	1.3	0.4	2.8	–	–	–	20.1	1.1	–	2.1	5.6	–	
MT5	1.5	0.4	2.2	1.6	0.1	0.9	2.8	18.0	3.0	3.1	34.1	25.1	3.4	1.7	4.8	11.7	7.2	2.9
AfriMT5	2.1	0.8	3.7	2.5	0.1	1.8	5.1	19.6	5.2	4.6	**35.0**	**26.7**	7.0	2.7	6.2	13.2	8.5	4.8
ByT5	9.5	1.8	5.5	3.8	0.1	6.0	8.3	21.8	12.1	8.4	30.1	24.4	14.7	6.0	7.5	14.0	10.9	8.4
AfriByT5	11.4	2.2	5.2	3.7	0.2	6.4	9.3	22.7	13.1	8.9	30.0	24.7	17.0	6.1	7.6	15.3	11.5	9.1
mBART50	18.6	2.4	5.3	6.2	0.8	9.7	8.9	21.1	12.0	10.0	34.1	25.8	16.8	7.5	10.0	**21.2**	13.2	10.0
AfriMBART	15.3	2.4	5.7	4.4	0.6	8.6	10.4	22.4	10.0	9.8	30.0	22.7	12.8	6.3	9.6	20.1	11.9	9.9
M2M-100	**22.7**	**2.9**	**6.4**	**7.1**	**1.0**	**12.4**	**16.0**	**24.7**	**14.3**	**11.5**	33.9	**26.7**	**24.7**	**8.8**	**12.8**	21.0	**15.4**	**13.6**
M2M-100-EN/FR	18.5	2.2	6.2	4.3	0.8	10.6	7.0	22.4	8.9	9.5	34.9	26.4	19.7	7.0	5.6	15.6	12.5	9.2
CHRF																		
M2M-100 0-shot	–	–	–	–	–	4.3	12.4	19.0	–	–	–	47.7	8.7	–	10.4	20.1	–	
MT5	10.0	7.4	9.7	11.5	7.9	9.1	23.6	41.1	24.9	21.6	64.1	53.7	22.8	17.8	20.8	36.0	23.9	21.2
AfriMT5	14.0	12.7	16.6	14.8	8.2	13.8	29.7	43.1	30.4	25.7	**64.7**	55.1	31.5	21.5	24.3	40.3	27.9	25.0
ByT5	27.8	17.7	23.8	16.1	8.8	22.9	31.3	46.5	40.0	32.2	58.1	52.5	38.6	27.9	25.5	40.3	31.9	29.6
AfriByT5	31.4	19.9	24.1	16.5	9.8	23.8	32.8	47.4	42.2	33.6	58.0	52.8	42.1	29.0	26.0	42.9	33.3	32.1
mBART50	42.3	22.0	27.7	25.7	16.0	31.9	32.6	45.9	41.1	36.7	64.2	54.4	43.0	35.6	31.1	50.2	37.5	36.2
AfriMBART	40.4	20.1	26.9	24.1	15.1	30.9	40.3	47.4	38.6	36.7	54.9	52.7	40.3	34.2	31.1	49.3	36.4	37.7
M2M-100	**48.2**	**23.1**	**30.9**	**27.6**	**16.7**	**35.7**	**43.3**	**50.0**	**45.5**	**39.0**	64.0	**56.4**	**52.0**	**38.2**	**35.9**	**51.2**	**41.1**	**41.2**
M2M-100-EN/FR	43.4	20.6	29.4	23.2	16.3	32.8	33.3	46.9	38.8	36.5	64.5	55.4	47.1	33.6	25.3	42.9	36.9	35.0

Table 3: **Results adding African Languages to Pre-Trained Models, en/fr-xx**. We calculate BLEU and CHRF on the news domain when training on only `NEWS` data from MAFAND-MT.

	xx-fr						*xx-en*											
Model	**bam**	**bbj**	**ewe**	**fon**	**mos**	**wol**	**hau**	**ibo**	**lug**	**luo**	**pcm**	**swa**	**tsn**	**twi**	**yor**	**zul**	**AVG**	**MED**
BLEU																		
M2M-100 0-shot	–	–	–	–	–	0.8	2.2	6.4	–	–	–	25.2	3.3	–	3.0	13.8	–	
MT5	2.5	0.9	1.1	2.4	0.7	1.3	5.8	18.9	12.6	6.4	42.2	29.5	9.5	4.6	12.3	22.4	10.8	6.1
AfriMT5	6.4	2.0	2.1	4.2	1.2	2.9	10.4	19.5	15.5	9.7	44.6	30.6	16.1	8.4	13.8	24.0	13.2	10.0
ByT5	10.0	2.7	4.1	4.9	1.5	7.2	12.9	21.0	19.8	12.1	39.4	27.1	18.6	9.8	11.5	22.8	14.1	11.8
AfriByT5	13.8	4.4	4.5	5.8	2.2	9.0	13.5	20.7	**21.1**	12.5	39.5	27.0	19.7	10.5	11.9	24.0	15.0	13.0
mBART50	6.8	0.3	1.7	0.8	0.6	6.3	11.5	13.2	14.5	9.1	44.2	29.0	2.0	0.5	8.1	31.1	11.2	7.4
AfriMBART	8.1	2.3	3.0	4.5	1.7	3.2	10.2	15.5	13.1	8.0	43.7	29.2	7.2	6.5	9.5	33.0	12.4	8.0
M2M-100	**22.1**	**5.4**	6.9	8.4	**2.8**	10.3	**17.0**	19.0	20.0	13.0	43.8	29.8	20.0	10.9	**16.0**	**37.8**	**17.7**	**16.5**
M2M-100-EN/FR	**22.1**	5.1	**7.4**	**9.1**	2.1	**10.5**	11.4	20.3	19.8	**14.0**	**45.2**	**30.0**	**21.4**	**11.7**	13.4	9.5	15.8	12.6
CHRF																		
M2M-100 0-shot	–	–	–	–	–	12.3	23.7	29.7	–	–	–	51.6	21.1	–	18.3	35.7	–	
MT5	19.4	15.1	17.0	17.9	10.9	16.2	26.3	43.5	36.3	26.1	66.9	53.7	32.2	25.2	31.1	43.9	30.1	26.2
AfriMT5	27.7	19.6	21.1	21.4	13.2	21.6	32.5	44.9	40.2	32.2	68.4	54.5	39.6	31.2	33.9	45.9	34.2	32.4
ByT5	31.2	21.8	24.8	20.5	15.4	26.2	33.2	46.4	45.4	34.1	62.0	50.6	42.4	32.9	31.4	42.5	35.0	33.0
AfriByT5	34.8	25.5	24.9	22.0	16.2	29.3	33.9	46.4	**47.1**	35.0	62.1	50.5	43.4	33.4	32.0	43.7	36.3	34.3
mBART50	26.0	17.1	20.9	20.2	17.1	26.6	32.0	37.9	39.0	31.0	68.2	53.5	20.1	19.4	26.7	49.0	31.5	26.6
AfriMBART	31.4	22.9	27.2	26.3	17.0	25.0	34.3	42.0	40.4	29.8	67.8	53.5	31.4	30.6	30.0	51.7	35.1	31.0
M2M-100	**45.9**	26.5	30.9	27.5	**17.7**	33.8	**38.7**	46.1	46.4	36.7	68.6	54.8	45.2	35.1	**38.1**	**55.5**	**40.5**	**38.4**
M2M-100-EN/FR	45.6	**26.9**	**32.2**	**28.7**	17.0	**34.3**	35.1	**46.6**	46.0	**37.6**	**69.0**	**55.0**	**46.3**	**36.0**	35.2	31.5	38.9	35.6

Table 4: **Results adding African Languages to Pre-Trained Models, xx-en/fr**. We calculate BLEU and CHRF on the news domain when training on only `NEWS` data from MAFAND-MT.

PROGRESS: HOW MUCH DATA WE MINIMALLY NEED TO PERFORM WELL IN TRANSFER LEARNING SETTING

A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation

Domain Shift Analysis: Is a small in-domain set essential for finetuning?

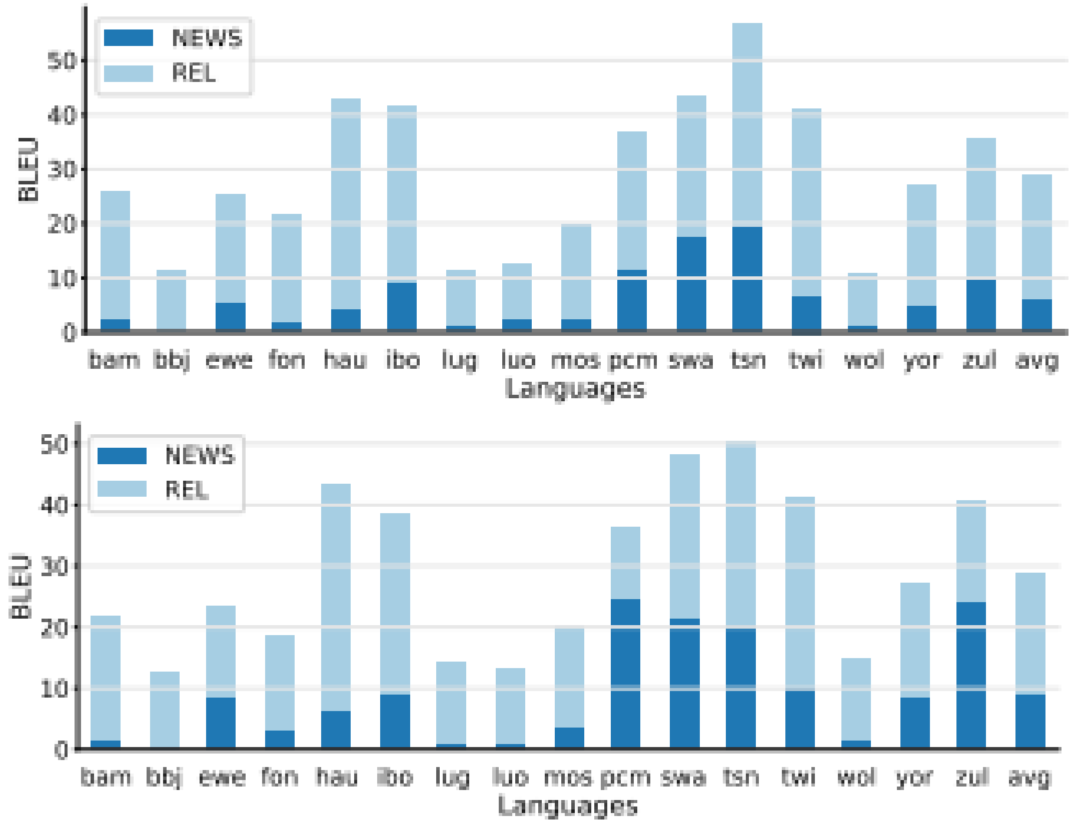


Figure 1: **Domain shift** of M2M-100 Transformer models trained on en/fr-xx (top) or xx-en/fr (bottom) REL domain and tested on the NEWS vs. REL domains.

If we train models only on previously available religious data, they are not capable of translating news well due to the strong domain bias. All models perform much worse on NEWS than on the REL domain

PROGRESS: HOW MUCH DATA WE MINIMALLY NEED TO PERFORM WELL IN TRANSFER LEARNING SETTING

So how well do we do when we adapt to domain shift & how much data do we need in the target domain to do "well"

PROGRESS: HOW MUCH DATA WE MINIMALLY NEED TO PERFORM WELL IN TRANSFER LEARNING SETTING

A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation

Domain Shift Adaptation Results: DS Adaptation HELPS even if sometimes very marginally

	xx-fr						*xx-en*											
Model	**bam**	**bbj**	**ewe**	**fon**	**mos**	**wol**	**hau**	**ibo**	**lug**	**luo**	**pcm**	**swa**	**tsn**	**twi**	**yor**	**zul**	**AVG**	**MED**
BLEU																		
Transformer																		
REL+NEWS	4.9	0.6	6.3	2.2	3.7	2.2	11.2	17.4	5.6	3.1	19.5	28.0	23.9	9.8	12.0	27.3	11.1	8.0
REL→NEWS	4.7	0.8	6.5	2.4	3.1	2.5	11.0	17.4	6.3	1.8	19.0	27.9	24.6	10.1	11.0	28.5	11.1	8.3
REL+NEWS→NEWS	5.8	1.0	7.1	2.4	**4.1**	2.6	13.2	18.2	6.8	3.7	21.4	28.7	24.5	10.4	12.6	30.1	12.0	8.8
M2M-100																		
REL+NEWS	24.0	5.8	10.9	9.7	2.3	10.1	15.3	21.1	21.1	13.3	**44.6**	29.4	27.0	12.5	17.4	30.6	18.4	16.4
REL→NEWS	20.3	5.9	11.4	9.6	2.3	10.5	17.4	**21.9**	20.6	13.7	44.3	**30.6**	27.7	**13.2**	**18.0**	36.0	19.0	17.7
REL+NEWS→NEWS	**25.8**	**6.3**	**11.6**	**9.9**	2.6	**11.5**	**18.2**	21.5	**22.4**	**14.3**	44.0	30.5	**27.8**	**13.2**	**18.0**	**38.1**	**19.7**	**18.1**
CHRF																		
Transformer																		
REL+NEWS	24.7	12.6	29.4	16.1	**17.6**	19.9	31.7	43.1	26.9	23.0	47.8	53.5	49.8	34.4	33.4	49.6	32.1	30.6
REL→NEWS	23.0	12.7	29.8	16.6	17.2	18.3	30.6	42.8	28.7	20.0	47.3	53.3	50.8	34.4	32.2	50.4	31.8	30.2
REL+NEWS→NEWS	26.5	14.7	30.7	17.6	18.8	21.8	33.8	44.0	29.5	24.7	50.8	54.1	50.6	35.1	34.4	51.4	33.7	32.2
M2M-100																		
REL+NEWS	47.1	27.5	36.4	27.9	16.6	34.0	36.8	47.5	47.2	37.3	**68.9**	54.7	53.0	38.4	40.2	53.3	41.7	39.3
REL→NEWS	44.5	27.7	37.0	28.2	16.8	34.4	39.6	**48.0**	47.0	38.0	68.7	**55.8**	53.6	**38.7**	40.7	56.4	42.2	40.2
REL+NEWS→NEWS	**49.0**	**28.5**	**37.2**	**28.9**	17.2	**35.3**	**40.2**	47.9	**48.5**	**38.3**	68.6	55.7	**54.0**	**38.7**	**41.0**	**57.7**	**42.9**	**40.6**

Table 6: **Results adapting to Domain Shift, xx-en/fr**. We calculate BLEU and ChrF on the news domain when training on different combinations of REL and NEWS.

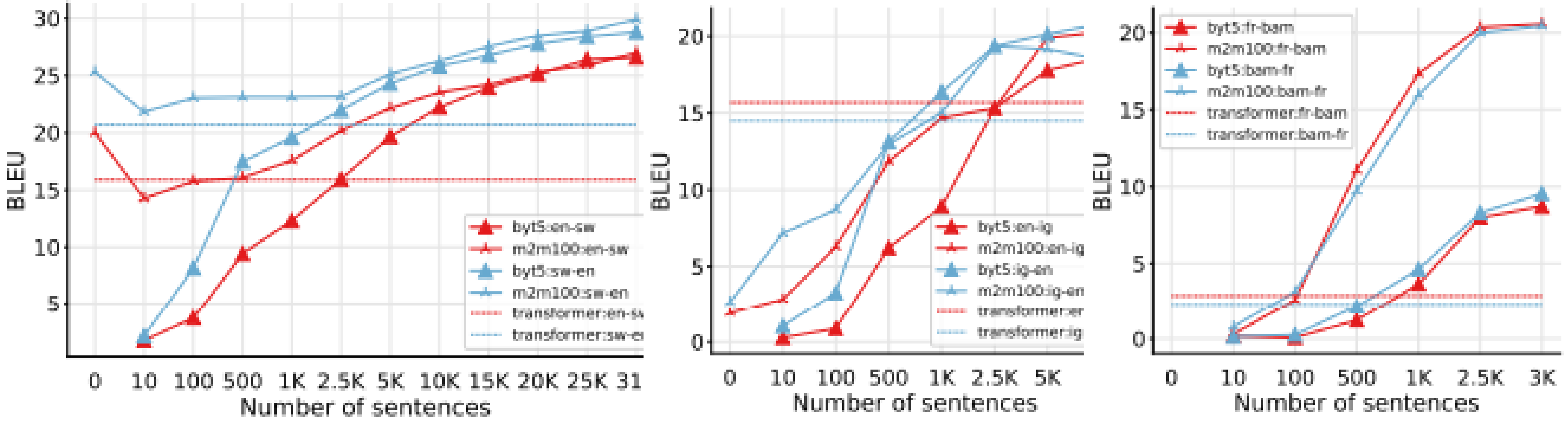
	fr-xx						*en-xx*											
Model	**bam**	**bbj**	**ewe**	**fon**	**mos**	**wol**	**hau**	**ibo**	**lug**	**luo**	**pcm**	**swa**	**tsn**	**twi**	**yor**	**zul**	**AVG**	**MED**
BLEU																		
Transformer																		
REL+NEWS	7.3	0.1	6.2	2.9	2.1	3.1	10.7	22.4	4.6	3.7	11.7	26.2	28.1	8.7	9.7	16.5	10.2	8.0
REL→NEWS	5.1	0.2	5.4	2.8	1.7	2.3	11.7	22.7	3.9	3.3	11.9	26.3	29.7	8.7	8.4	20.3	10.3	6.9
REL+NEWS→NEWS	8.5	0.3	6.5	3.2	**2.2**	3.7	12.0	23.6	5.1	4.3	13.8	26.6	29.3	9.0	9.7	20.1	11.1	8.8
M2M-100																		
REL+NEWS	23.0	2.8	7.7	6.5	0.9	11.2	12.9	24.7	13.9	11.6	**35.1**	23.3	29.0	9.7	12.4	18.3	15.2	12.6
REL→NEWS	20.3	**3.1**	7.7	**7.5**	1.1	12.0	15.0	**26.0**	15.4	11.9	35.0	**27.7**	**31.9**	10.0	13.4	**22.9**	16.3	14.2
REL+NEWS→NEWS	**24.7**	**3.1**	**8.9**	7.4	1.1	**12.7**	**15.9**	25.8	**15.7**	**12.0**	34.2	27.3	**31.9**	**10.2**	**13.9**	22.6	**16.7**	**14.8**
CHRF																		
Transformer																		
REL+NEWS	25.6	9.6	30.6	14.5	17.7	18.9	36.7	46.7	30.5	26.4	37.8	55.3	55.0	36.7	30.6	50.0	32.7	30.6
REL→NEWS	18.2	11.2	27.1	15.4	18.3	15.9	37.4	47.2	28.7	24.4	38.3	55.5	56.3	36.6	28.9	53.0	32.0	28.8
REL+NEWS→NEWS	27.4	12.8	31.5	16.5	19.9	20.2	38.3	48.3	30.6	27.7	42.6	55.6	56.3	37.7	30.6	53.4	34.3	31.0
M2M-100																		
REL+NEWS	46.8	22.1	36.7	26.2	16.0	33.5	38.4	50.1	44.5	38.1	64.7	53.0	57.2	39.7	35.2	53.1	41.0	39.0
REL→NEWS	44.1	22.6	34.1	27.7	16.8	34.7	41.3	51.3	45.6	38.6	**64.7**	**57.2**	59.3	40.6	37.1	**56.3**	42.0	41.0
REL+NEWS→NEWS	**49.9**	**23.5**	**37.5**	**28.5**	**16.8**	**35.8**	**42.1**	**51.3**	**46.9**	**39.4**	64.2	57.0	**59.5**	**40.8**	**37.4**	**56.3**	**42.9**	**41.4**

Table 5: **Results adapting to Domain Shift, en/fr-xx**. We calculate BLEU and ChrF on the news domain when training on different combinations of REL and NEWS.

PROGRESS: HOW MUCH DATA WE MINIMALLY NEED TO PERFORM WELL IN TRANSFER LEARNING SETTING

A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation

Even small **good & high quality** in the target domain help.



(a) eng-swa (b) eng-ibo (c) fr-bam

Figure 2: **Number of fine-tuning sentences** needed to exceed the performance of a bilingual Transformer model.

PROGRESS: HOW MUCH DATA WE MINIMALLY NEED TO PERFORM WELL IN TRANSFER LEARNING SETTING

What if we want to start training from scratch? How do we cope with data scarcity, increase model robustness & ensure generalization ?

PROGRESS: ACTIVE LEARNING FOR AFRICAN NLP DOWNSTREAM TASKS

AfroLM: A Self-Active Learning-based Multilingual Pre-trained Language Model for 23 African Languages

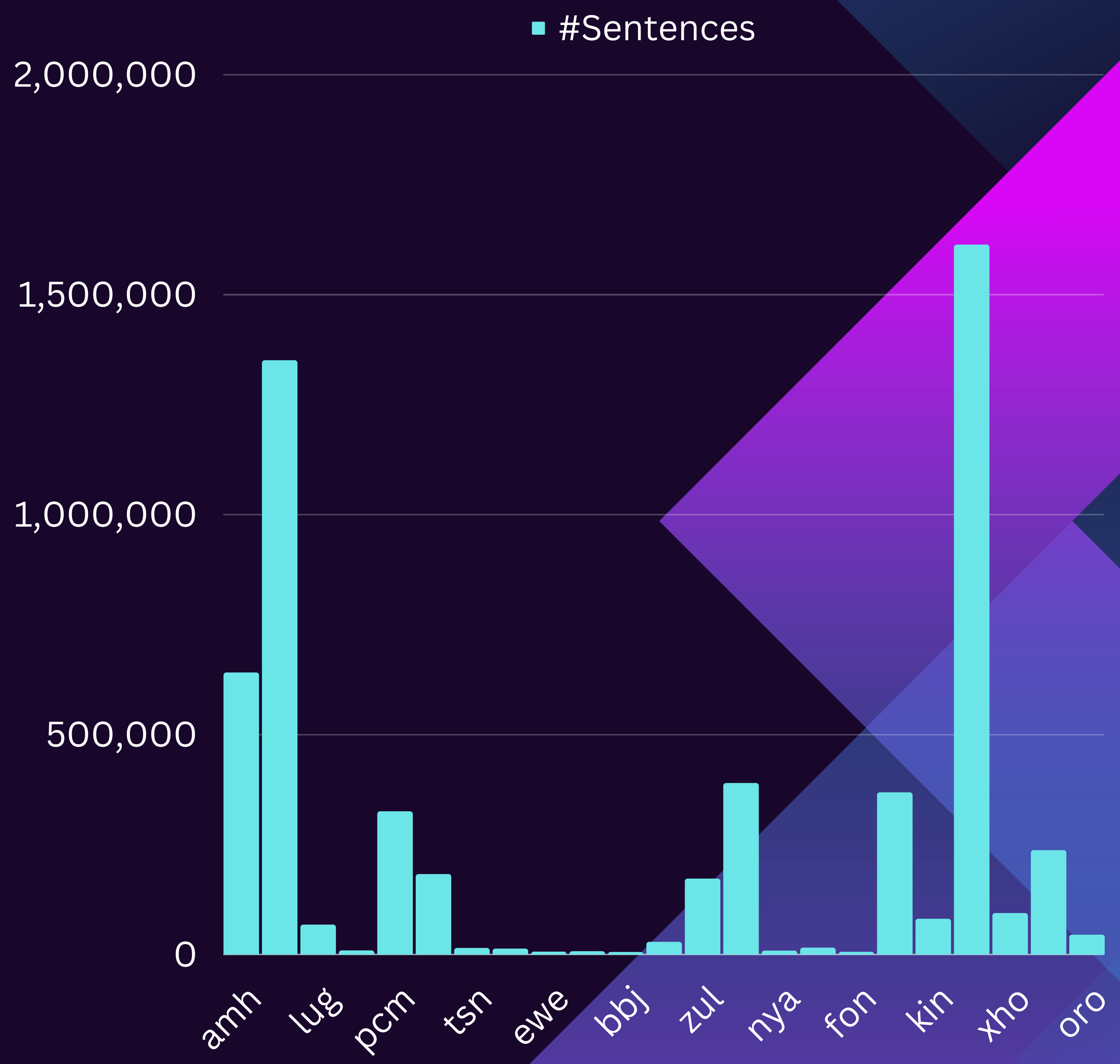
EMNLP 2022





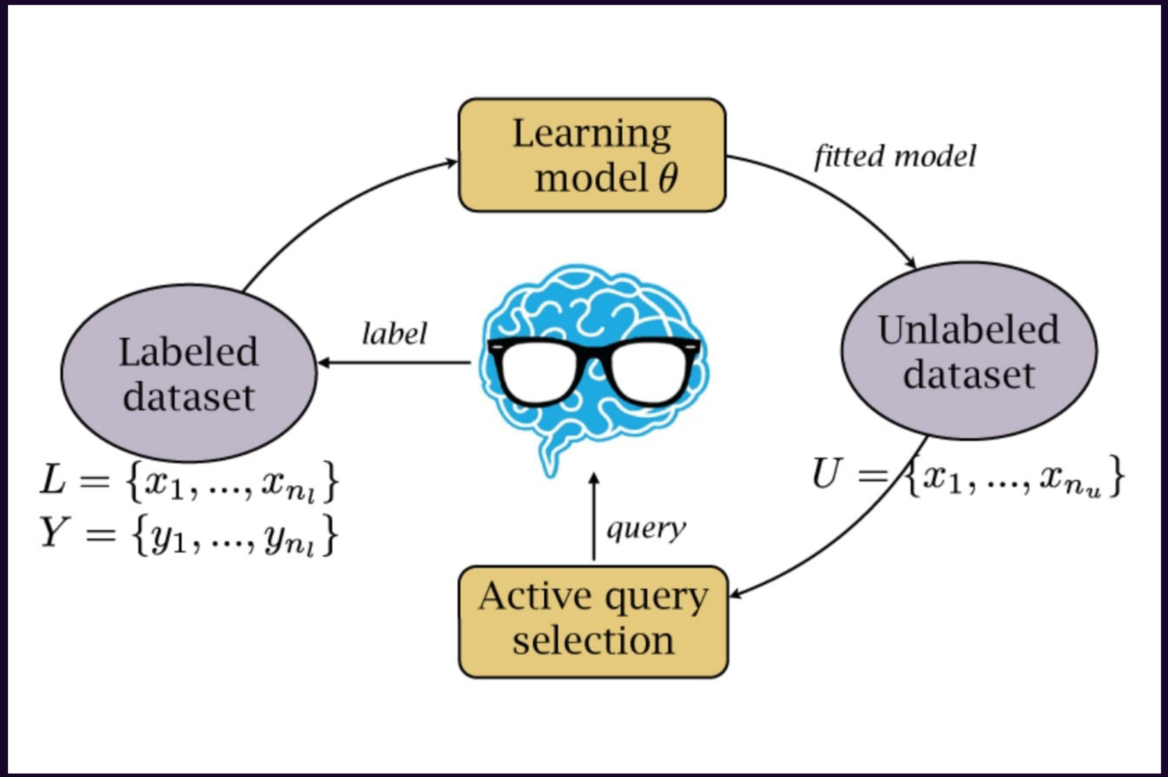

PROGRESS: AFROLM

Some Statistics



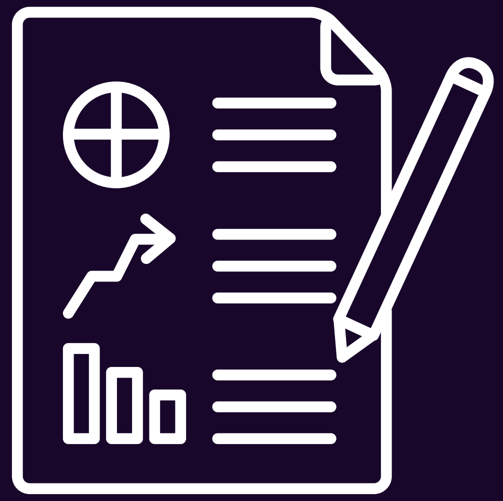
AfroLM dataset's # of sentences per language

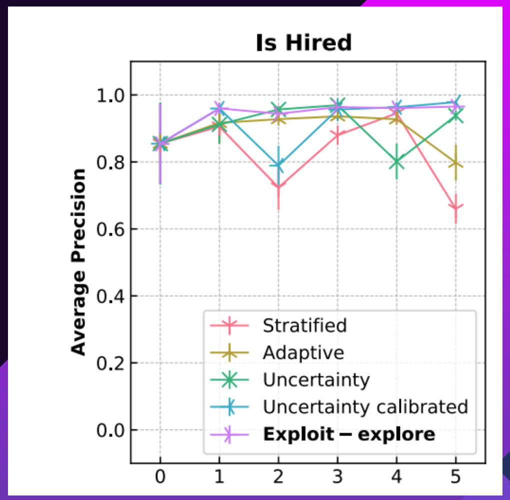
PROGRESS: AFROLM --- ACTIVE LEARNING & BENEFITS



Active learning is a form of semi-supervised learning algorithm where the learner learns to choose which data to learn from. The learner does this by actively querying an authority source (called oracle) to learn the correct prediction for a given problem.

The goal of this iterative learning approach is to speed along the learning process, especially when there is a lack of a large (huge) labeled dataset to practice traditional supervised learning methods.





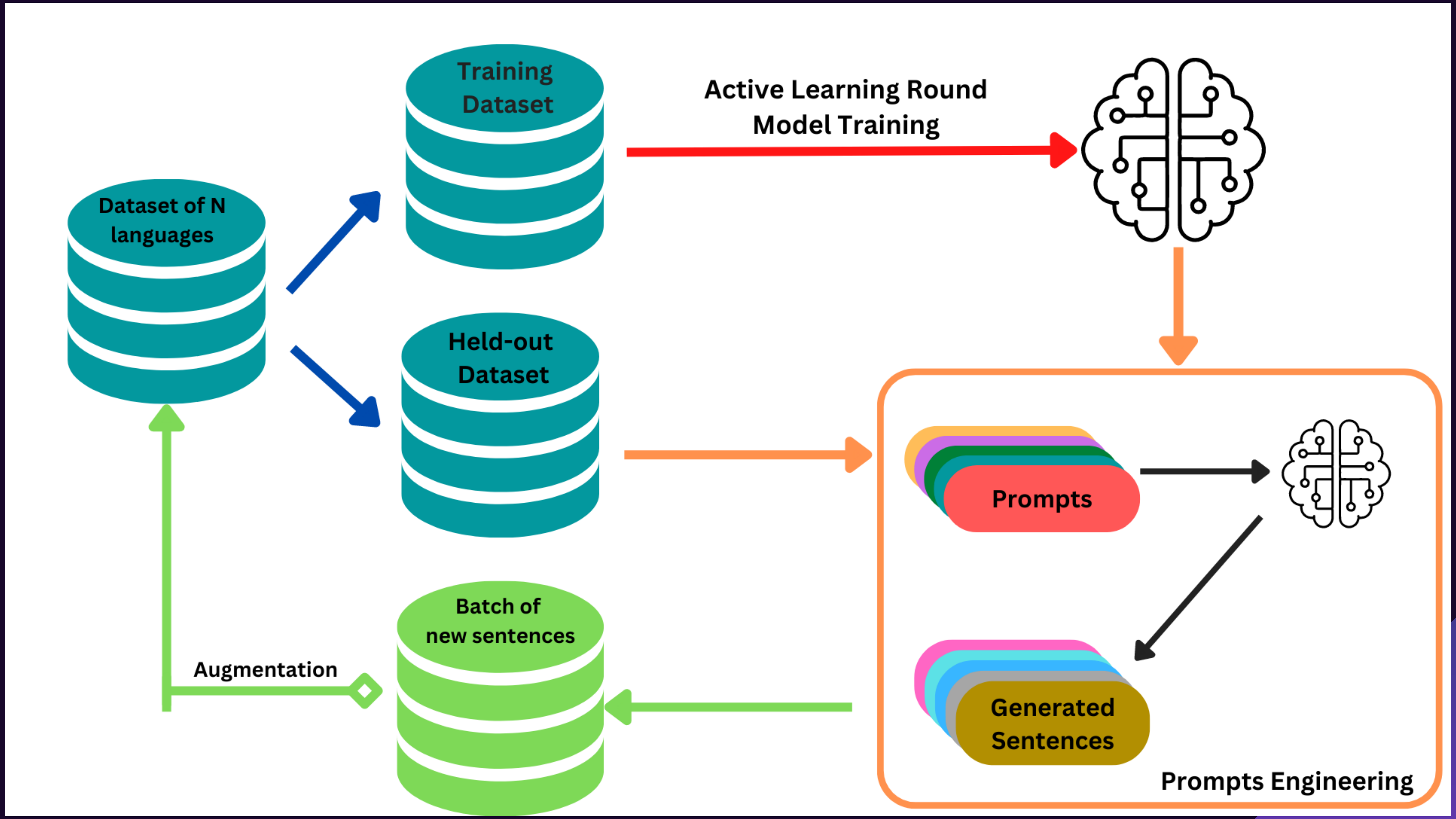
WITH ONLY 20% OF LABELLED SAMPLES

ACHIEVED 98-99% ACCURACY

Very Crucial in AfricaNLP & Low-resource context

Other approaches need 50% of data

Our Self-Active Framework



Experiments

- Named Entity Recognition (NER)
  - MasakhaNER (**10 African Languages**, TACL 2022 & ACL 2022)
  - MasakhaNER 2.0 (**20 African Languages**, EMNLP 2022)
- Text Classification
  - Hausa and Yorùbá news text classification dataset from (Hedderich et al., 2020)
- Sentiment Analysis (OOD Experiments)
  - Movies Domain
  - Twitter Domain → Movies Domain

More details about hyperparameters in our paper

Results and Discussion

- MasakhaNER (10 African Languages)

Language	AfriBERTa-Large	AfroLM-Large (w/o AL)	AfroLM-Large (w/ AL)	mBERT	XLMR-base
amh	73.82	43.78	73.84	00.00	70.96
hau	90.17	84.14	91.09	87.34	87.44
ibo	87.38	80.24	**87.65**	85.11	84.51
kin	73.78	67.56	72.84	70.98	73.93
lug	78.85	72.94	80.38	80.56	80.71
luo	70.23	57.03	**75.60**	72.65	75.14
pcm	85.70	73.23	87.05	87.78	87.39
swa	87.96	74.89	87.67	86.37	87.55
wol	61.81	53.58	65.80	66.10	64.38
yor	81.32	73.23	79.37	78.64	77.58
avg	79.10	68.06	*80.13*	71.55	79.16
avg (excl. amh)	79.69	70.76	*80.83*	79.50	80.07

mBERT, and XLMR are trained on >= ~2.5TB of data, AfriBERTa was trained on ~ 0.93 GB, and AfroLM was trained on ~0.73GB data

Results and Discussion

- MasakhaNER 2.0 (11 additional African Languages)

Model	bam	bbj	ewe	fon	mos	nya	sna	tsn	twi	xho	zul	AVG
MPLMs pre-trained on from scratch on African Languages												
AfriBERTa-Large	78.60	71.00	86.90	79.90	71.40	88.60	92.40	83.20	75.70	85.00	81.70	81.31
AfroLM-Large (w/ AL)	**80.40**	**72.91**	**88.14**	**80.48**	**72.14**	**90.25**	**94.46**	**85.38**	**77.89**	**87.50**	**86.31**	**83.26**
MPLMs adapted to African Languages												
mBERT	78.90	60.60	86.90	79.90	71.40	88.60	92.40	86.40	75.70	85.00	81.70	80.68
XLMR-base	78.70	72.30	88.50	81.90	72.70	89.90	93.60	86.10	78.70	87.00	84.60	83.09

Results and Discussion

- Text Classification and Sentiment Analysis (OOD Experiments)

Language	In AfriBERTa?	In AfroLM?	AfriBERTa-Large	AfroLM-Large (w/o AL)	AfroLM-Large (w/ AL)
hau	✓	✓	90.86	85.57	**91.00**
yor	✓	✓	**83.22**	75.30	82.90

Results and Discussion

- Text Classification and Sentiment Analysis (OOD Experiments)

Language	In AfriBERTa?	In AfroLM?	AfriBERTa-Large	AfroLM-Large (w/o AL)	AfroLM-Large (w/ AL)
hau	✓	✓	90.86	85.57	**91.00**
yor	✓	✓	**83.22**	75.30	82.90

AfroLM generalized better in OOD settings

Models	Yoruba F1-score
AfroLM-Large (w/o AL)	
Movies	83.12
Twitter → Movies	41.28
AfroLM-Large (w/ AL)	
Movies	**85.40**
Twitter → Movies	**68.70**
AfriBERTa-Large	
Movies	82.70
Twitter → Movies	65.90

Table 7: **Out-Of-Domain Sentiment Analysis Performance:** F1-scores on YOSM test set after 20 epochs averaged over 5 seeds.

AfroLM

Overall Conclusion

- (self-)Active Learning is very data efficient and high-performing
- AfroLM achieves SOTA against AfriBERTa, mBERT, and XLMR on NER, Text Classification, and Sentiment Analysis tasks
- AfroLM is generalizes better in across various domains

We can build powerful AI models, yet data-centric and very efficient

ADAPTING PRETRAINED ASR MODELS TO LOW-RESOURCE CLINICAL SPEECH USING EPISTEMIC UNCERTAINTY-BASED DATA SELECTION

CONTEXTUALIZATION

Assume we want to adapt an ASR model to a set of new and diverse languages

CONTEXTUALIZATION

The languages are very low-resourced:

- Very limited labeled data
- High morphological complexity
- Maybe some/lots of languages unlabeled data but no budget to label them because human labor is expensive

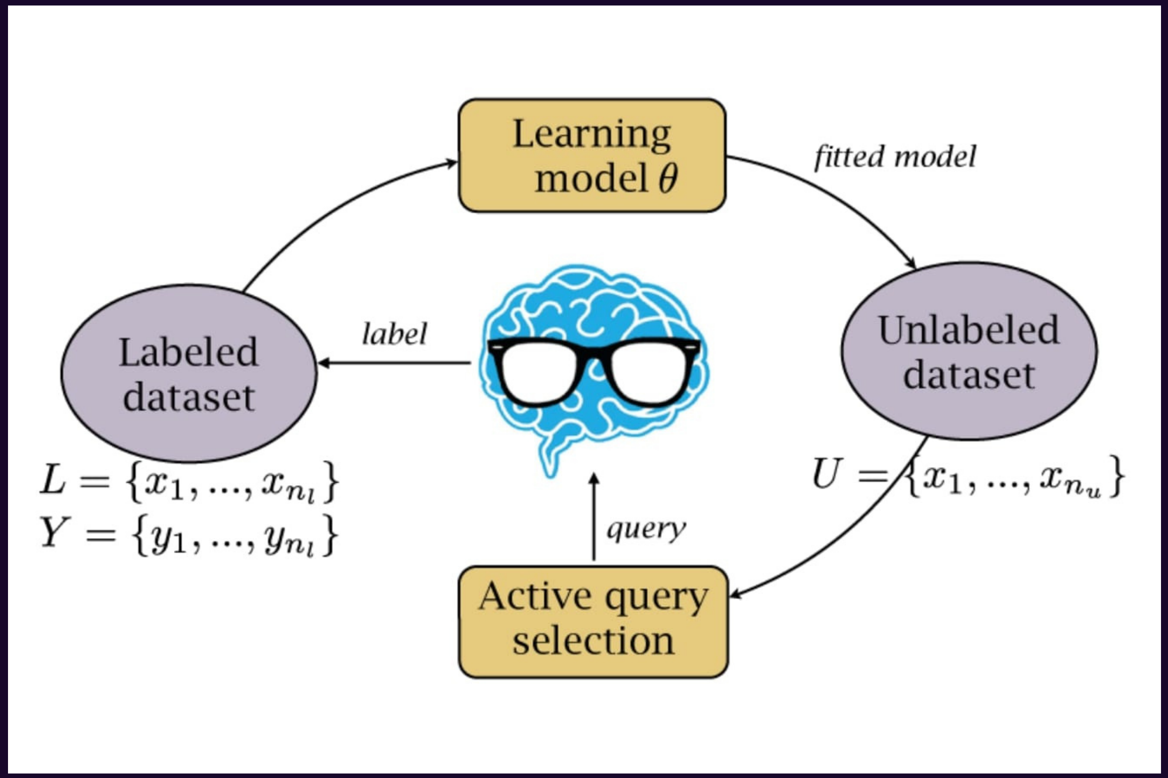
PROBLEM STATEMENT: MAIN QUESTIONS & GOALS

- How do we cope with data scarcity i.e. how to use efficiently use the small data available efficiently while maximizing the downstream performance on EACH language and domain?

- How do we reduce the cost of annotation while ensuring high-quality labeling?

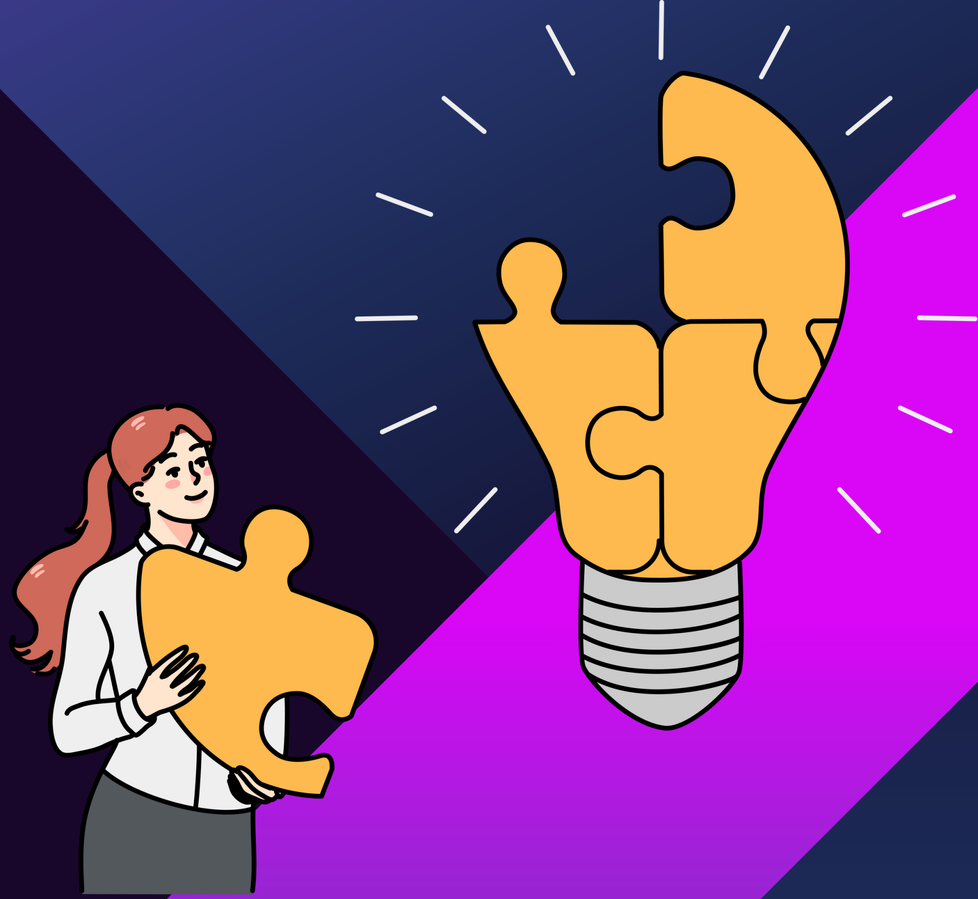
- How to also increase model robustness & and ensure generalization?

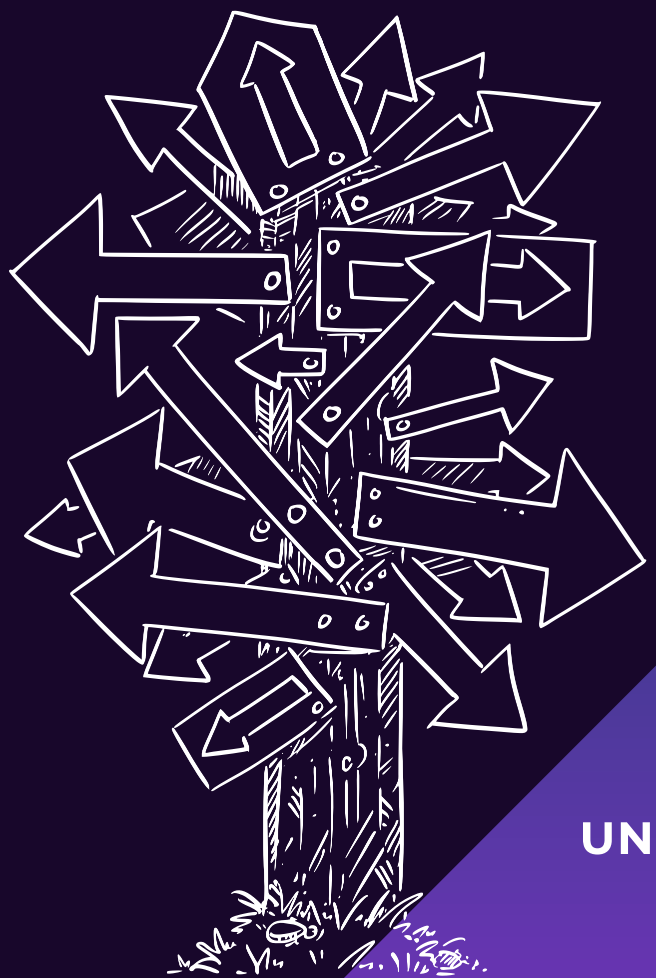
SOLUTION: ACTIVE LEARNING & UNCERTAINTY QUANTIFICATION



ACTIVE LEARNING

+





UNCERTAINTY QUANTIFICATION

Epistemic Uncertainty

Epistemic uncertainty refers to uncertainty caused by a lack of knowledge. But the good news is that it can in principle be reduced based on additional information.

Example of the lack of "uncertainty awareness": EfficientNet predictions (Tan and Le, 2019) on test images from ImageNet:

For the left image, the neural network predicts "typewriter keyboard" with 83.14% certainty, and for the right image "stone wall" with 87.63% certainty.

Machine learning is inseparable from uncertainty

The Epistemic Uncertainty can be defined as the variance of the model prediction

$$
\begin{aligned}
V(g(x,\theta)) &= \mathbb{E}_{\theta_t \sim q}[g(x,\theta_t)^2] - (\mathbb{E}_{\theta_t \sim q}[g(x,\theta_t)])^2 \\
&= \frac{1}{T}\sum_{i=1}^{T} f(x,\theta_t)^2 - (\frac{1}{T}\sum_{i=1}^{T} f(x,\theta_t))^2
\end{aligned}
$$

The Epistemic Uncertainty can be defined as the variance of the model prediction

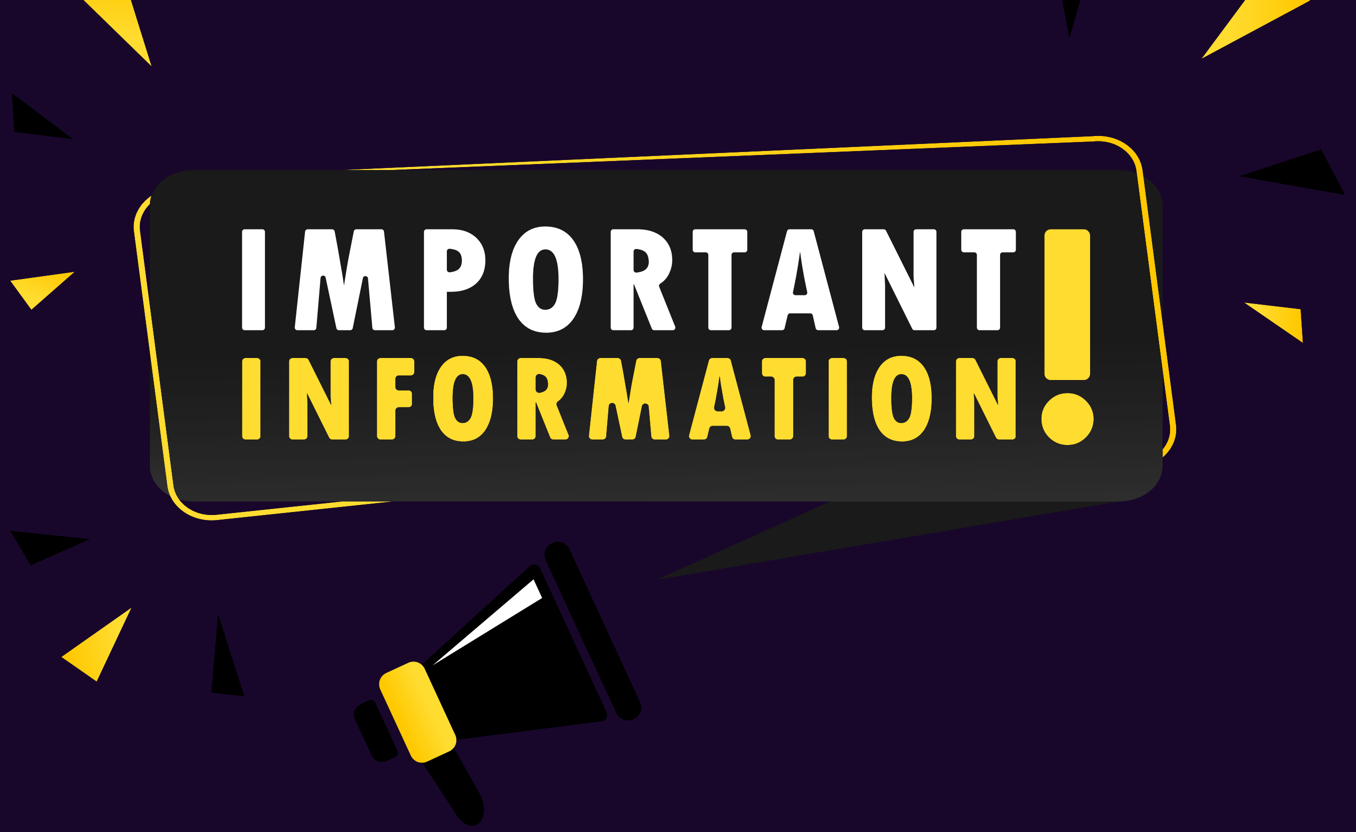
$$V(g(x,\theta)) = \mathbb{E}_{\theta_t \sim q}[g(x,\theta_t)^2] - (\mathbb{E}_{\theta_t \sim q}[g(x,\theta_t)])^2$$
$$= \boxed{\frac{1}{T}\sum_{i=1}^{T} f(x,\theta_t)^2 - (\frac{1}{T}\sum_{i=1}^{T} f(x,\theta_t))^2}$$

Quantity to reduce: Reducing the variance would imply more knowledge of the model about the data, and therefore more reliability

But how to reduce this quantity?

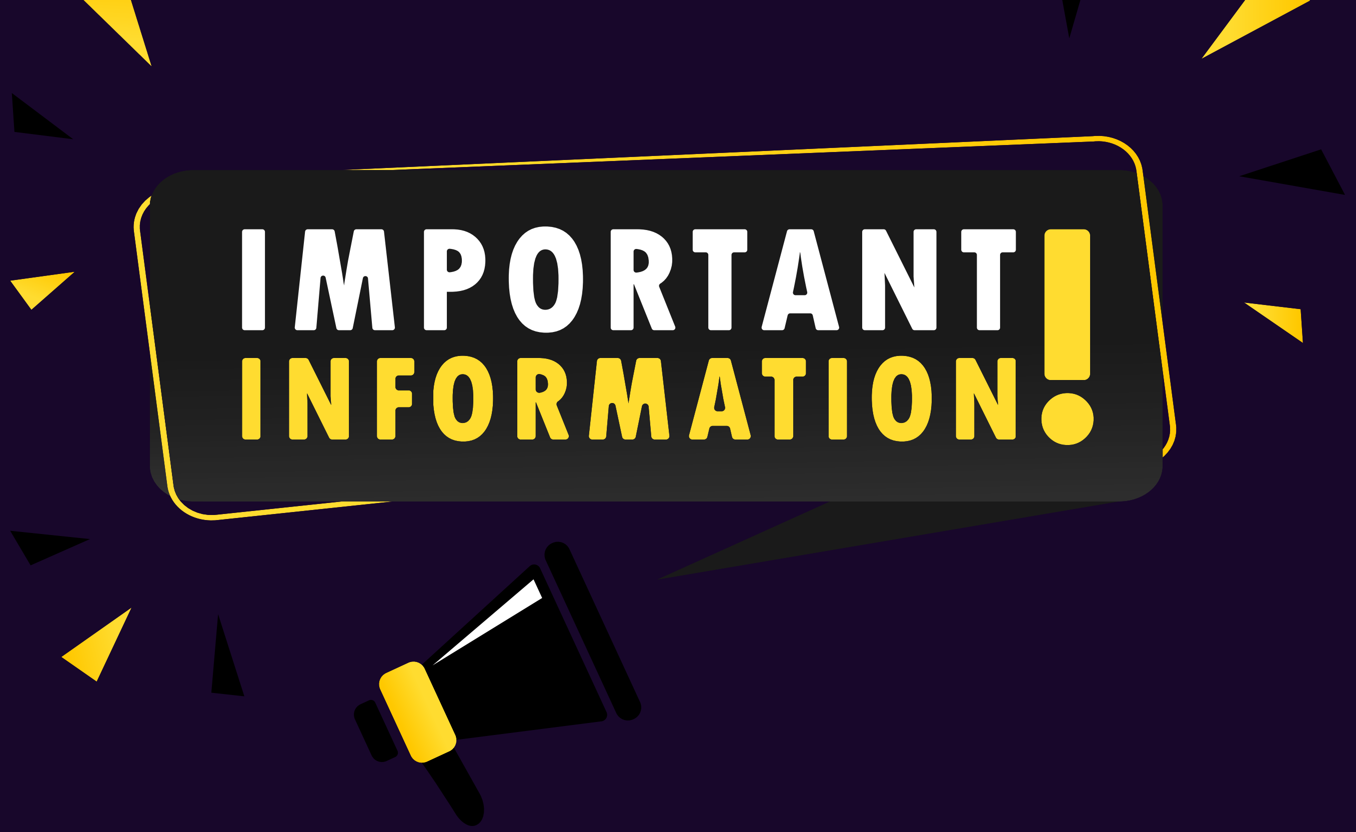
$$
\begin{aligned}
V(g(x,\theta)) &= \mathbb{E}_{\theta_t \sim q}[g(x,\theta_t)^2] - (\mathbb{E}_{\theta_t \sim q}[g(x,\theta_t)])^2 \\
&= \boxed{\frac{1}{T}\sum_{i=1}^{T} f(x,\theta_t)^2 - (\frac{1}{T}\sum_{i=1}^{T} f(x,\theta_t))^2}
\end{aligned}
$$

Mutual



Mutual information between two entities tells us to what extent knowledge of one entity reduces uncertainty about the other entity

Mutual



$$I(Y,H) = \mathbf{E}_{p(y,h)}\left\{\log_2\left(\frac{p(y,h)}{p(y)p(h)}\right)\right\}$$

$$\begin{aligned}\mathbb{I}[y,\omega|\mathbf{x},\mathcal{D}_{\text{train}}] &:= \mathbb{H}[y|\mathbf{x},\mathcal{D}_{\text{train}}] - \mathbb{E}_{p(\omega|\mathcal{D}_{\text{train}})}\left[\mathbb{H}[y|\mathbf{x},\omega]\right] \\ &= -\sum_c p(y=c|\mathbf{x},\mathcal{D}_{\text{train}}) \log p(y=c|\mathbf{x},\mathcal{D}_{\text{train}}) \\ &\quad + \mathbb{E}_{p(\omega|\mathcal{D}_{\text{train}})}\left[\sum_c p(y=c|\mathbf{x},\omega) \log p(y=c|\mathbf{x},\omega)\right]\end{aligned}$$

If we learn an objective that maximizes the information obtained about the model parameters, that is, maximizes the mutual information between the predictions and the posterior model, then we reduce the uncertainty and improve the high-dimensional representation of the data

How to estimate epistemic uncertainty ?

There are many ways to estimate epistemic uncertainty, but the two most common methods use Bayesian neural networks

Monte Carlo (MC) Dropout and Deep Ensembles.

- Evaluation Metric: Word Error Rate (WER)
- Selection Criteria: Uncertainty WER (U-WER)
  - Computed using McDropout over the predicted speech transcriptions
    - MC-Dropout helps quantify the model uncertainty without sacrificing either computational complexity or test accuracy and can be used for all kinds of models trained with dropout.
  - Sampling Mode: Select top-**k** ***most uncertain*** samples from the pool, at round ***r***

DATASET: AFRISPEECH-200

AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR (TACL 2023)

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, **Bonaventure F.P. Dossou**
Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, Clinton Mbataku

- 200hrs of recordings
- 67577 audio clips
- 2463 unique speakers
- 120 languages and accents
- First clinical (+medical dictation) and general domain ASR Model

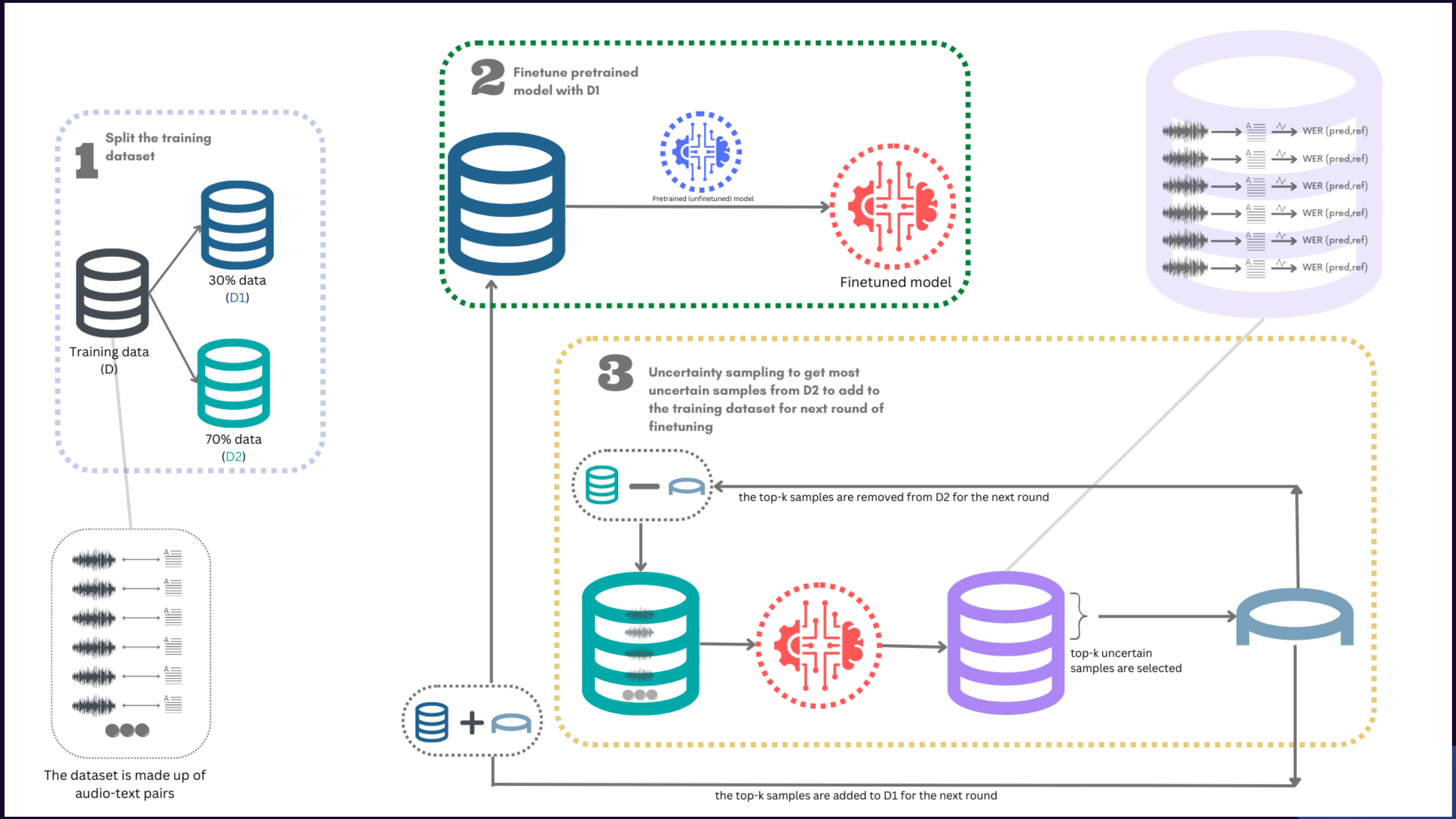
Speaker Gender Ratios - # Clip %	
Female	57.11%
Male	42.41%
Other/Unknown	0.48%
Speaker Age Groups - # Clips	
<18yrs	1,264 (1.87%)
19-25	36,728 (54.35%)
26-40	18,366 (27.18%)
41-55	10,374 (15.35%)
>56yrs	563 (0.83%)
Unknown	282 (0.42%)
Clip Domain - # Clips	
Clinical	41,765 (61.80%)
General	25,812 (38.20%)

Table 2: Dataset statistics.

Item	Train	Dev	Test
# Speakers	1466	247	750
# Hours	173.4	8.74	18.77
# Accents	71	45	108
Avg secs/speaker	425.80	127.32	90.08
clips/speaker	39.56	13.08	8.46
speakers/accent	20.65	5.49	6.94
secs/accent	8791.96	698.82	625.55
# general domain	21682	1407	2723
# clinical domain	36318	1824	3623

Table 3: Dataset splits showing speakers, number of clips, and speech duration in Train/Dev/Test splits.

OVERVIEW OF THE SYSTEM: INCREASING ROBUSTNESS OF PRETRAINED ASR MODELS BY INCOPORATING EPISTEMIC UNCERTAINTY



Work inspired by our previous work **AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages (Dossou et. al., EMNLP 2022)**

RESULTS: INCREASING ROBUSTNESS OF PRETRAINED ASR MODELS BY INCOPORATING EPISTEMIC UNCERTAINTY

Results of iterative epistemic uncertainty-based (uncertainty sampling) data selection

Model	General			Clinical			Both		
	Baseline	EU-Random	EU-Most	Baseline	EU-Random	EU-Most	Baseline	EU-Random	EU-Most
Wav2vec	0.4980	0.1111	**0.1011**	0.5610	0.3571	**0.2457**	0.5300	0.1666	**0.1266**
**Hubert	**0.1743**	—	0.1901	0.2907	—	**0.2594**	—	—	—-
**Nemo	0.2824	—	**0.1765**	0.2600	—	**0.2492**	—	—	—-

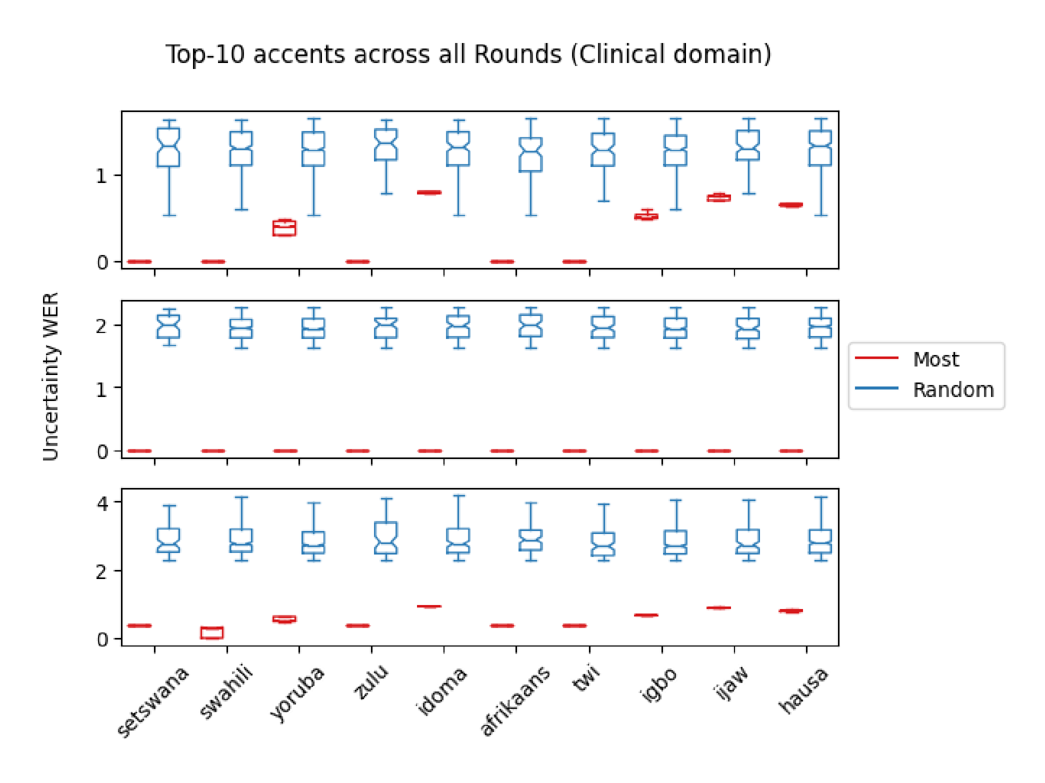
Dataset	Split and Size for our approach				Finetuning Epochs	Baseline (Entire training dataset)	Uncertainty Sampling w/ *most* (Train + αAug)
	Train	Aug	Top-k	Test			
SautiDB (Afonja et al., 2021a)	234	547	92	138	50	0.50	**0.12**
MedicalSpeech	1598	3730	1333	622	5	0.30	**0.28**
CommonVoices English (v10.0)	26614	62100	10350	232	5	0.50	**0.22**

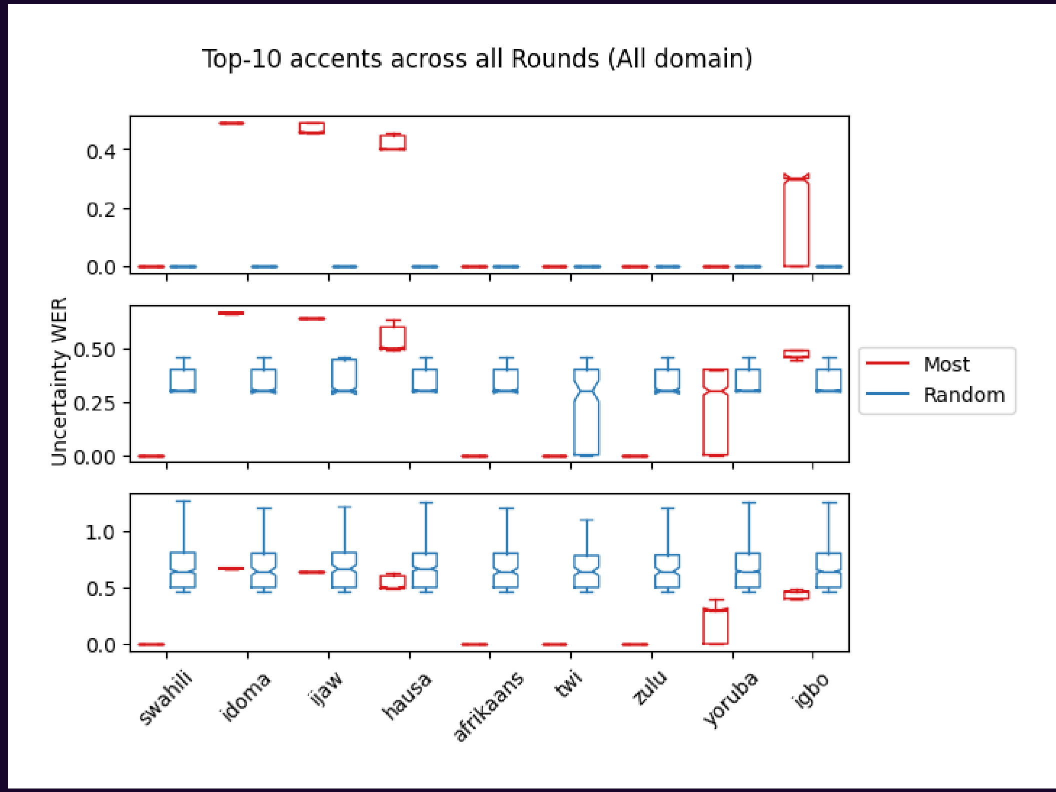
Outperforms all massively pretrained ASR models using ~40% less data

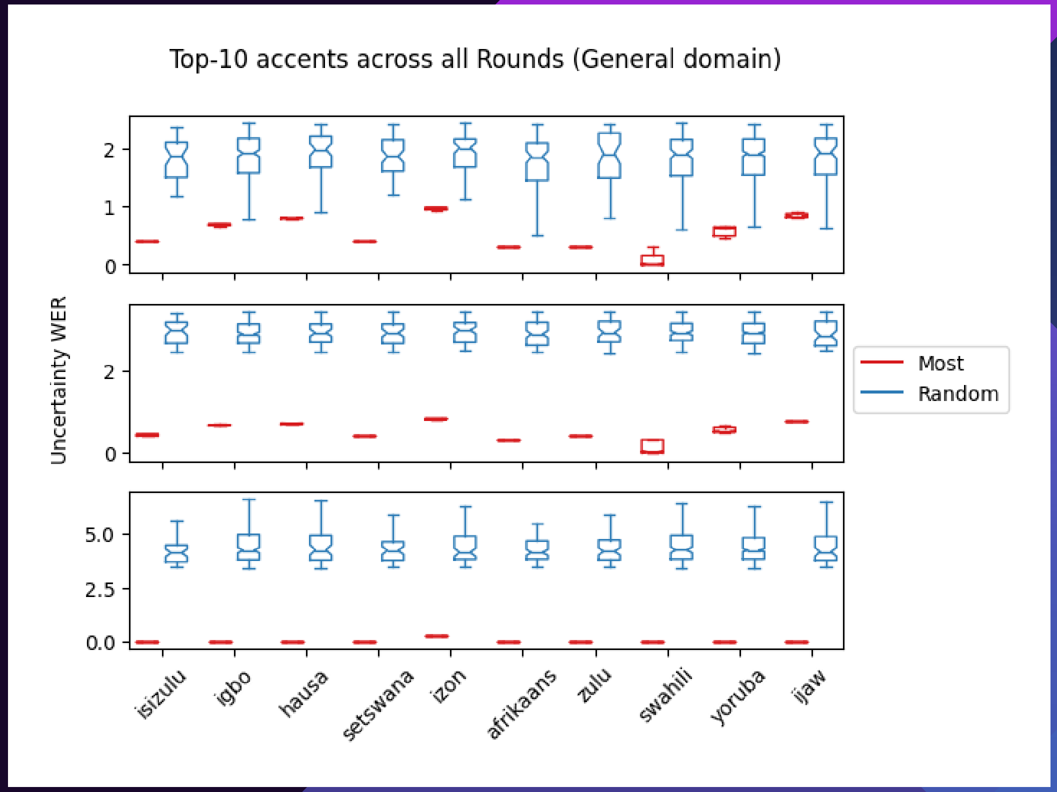
Model & Dataset agnostic: Performs well across several domains and datasets

INCREASING ROBUSTNESS OF PRETRAINED ASR MODELS BY INCOPORATING EPISTEMIC UNCERTAINTY

We defined an uncertainty word error rate (U-WER) that improved over active adaptation cycles.







Our approach is viable and efficient for building generalizable ASR models in the context of accentuated African Clinical ASR, where training datasets are really scarce.

Our analyses suggest that our approach enables ASR models to select and learn from the most informative data samples making it very suitable for low-resource settings.

CONCLUSION

- Our multi-round adaptive learning approach with uncertainty sampling is very data efficient and high-performing
- Our approach achieves SOTA (compared to w2v, Hubert, Nemo) on African Accents Transcription Task
- Better generalization: our approach is model, domain and dataset agnostic

We can build powerful AI models, yet data-centric and very efficient

FUTURE WORKS

- Explore trade-offs between adaptation rounds and the number of new data points selected at each round (query size)
- Improve Computational Complexity and Limitations
- Extend analyses to phoneme level for better explainability

The bigger Picture

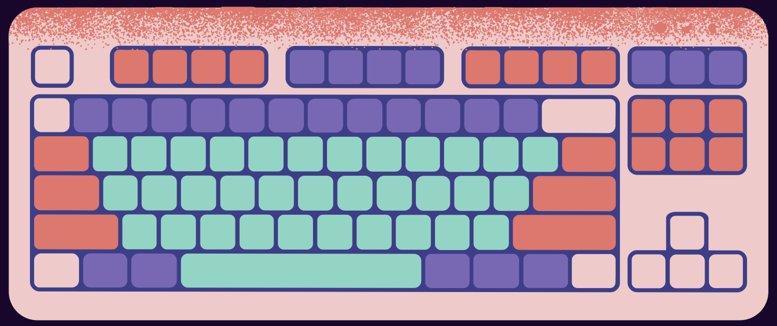
Technological (AI) Revolution
NOT TO MISS !!!!

- Focus on creating more resources for African Languages
- Engage in community efforts (e.g. Masakhane, GhanaNLP, etc.)
- Lead more Afro-centric research projects (more representation at top-tier NLP and AI conferences
- Build and Scale AI techniques proper to African Languages

My African Dream - Vision

Multilingual African Voice Assistants

African languages Translators, built by us for us



More keyboards and auto-complete systems

More research scholars, young people passionate by challenges and innovations

THANK YOU AND FEEL FREE TO REACH OUT

For More

Personal Website

Publications