

---

# Multilingual NLP and Cross-lingual Transfer

---

**David Ifeoluwa Adelani**

COMP-550

November 17, 2024

---

# Progress of language technology in English

How about other languages ? E.g. Yoruba

Great improvement in Language Models like **ChatGPT**.

**ChatGPT trained largely on English texts**

 You  
who brought Ebola disease to Nigeria?  
  
< 3 / 3 >

 ChatGPT  
The Ebola virus was brought to Nigeria by Patrick Sawyer, a Liberian-American, in July 2014. He arrived in Lagos, Nigeria's largest city, on a flight from Liberia, already infected with the virus. This marked the beginning of the Ebola virus outbreak in Nigeria during the West African Ebola epidemic. The rapid response of the Nigerian health authorities and international organizations helped to quickly contain the outbreak within the country.



 You  
tani o mu arun Ebola wa si Naijiria?  
  
< 2 / 2 >

**Answer retrieval is worse than English**

 ChatGPT  
Ko si iroyin tuntun kan ti o nka si ajakale-arun Ebola ti o de si Naijiria laipe. Alaye to wa ni lilo ju lo ni ti ajakale-arun Ebola ti o waye ni ḥodun 2014, eyi ti o je apakan kekere ti ajakale-arun Ebola ti o bère ni Guinea. Ajakale-arun naa je eyi ti o tan kaakiri orile-edé Naijiria ati pe o je igba akoko ti aisan yii de agbegbe lwo-orun Afirika. Ajakale-arun ti tèle ti ni opin si awon orile-edé ni Central Africa ni igba atijo [1].

< 2 / 2 >

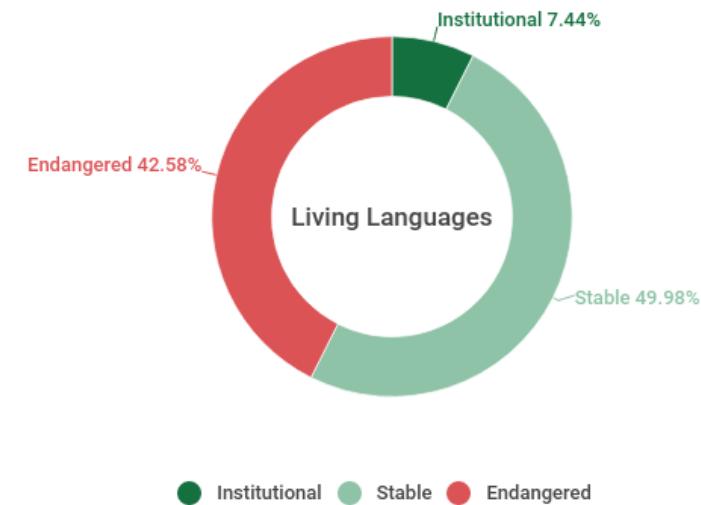
 You  
translate your answer to English

**Translation seem to be correct**

 ChatGPT  
There is no recent news regarding an Ebola outbreak arriving in Nigeria. The most frequently mentioned information relates to the Ebola outbreak that occurred in 2014, which was a minor part of the larger Ebola epidemic that began in Guinea. This outbreak spread to Nigeria, marking the first time the disease reached West Africa, having previously been confined to Central African countries in the past [1].

# Not many languages benefit from NLP

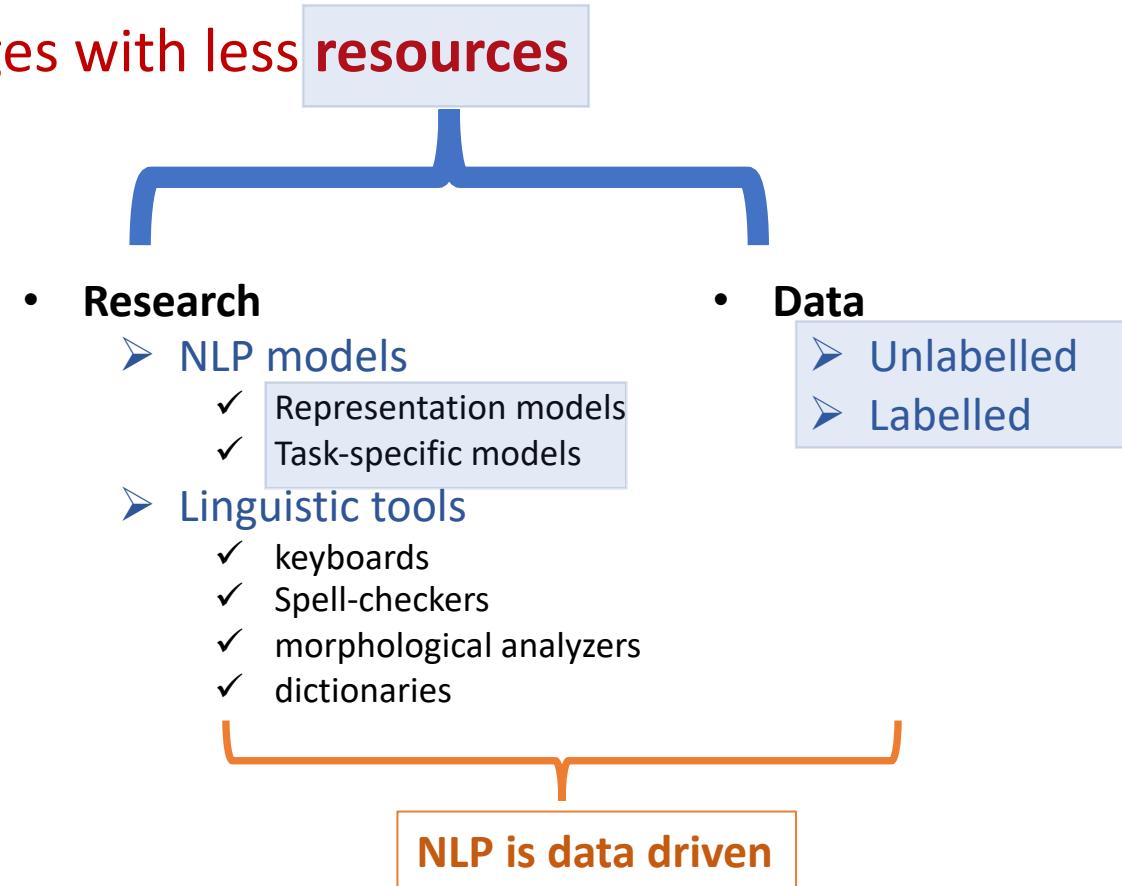
- 7,151 known languages (Ethnologue, 2022)
- Around 400 have more than 1M speakers
- Around 1,200 languages have more than 100k
- 2,000 have fewer than a 1,000 speakers
- Only 7.4% are institutional
  - Often used by governments, schools, and mass media
  - Even less are supported by language technologies.



Ethnologue

# Under-resourced languages

Under-resourced languages == **languages with less resources**

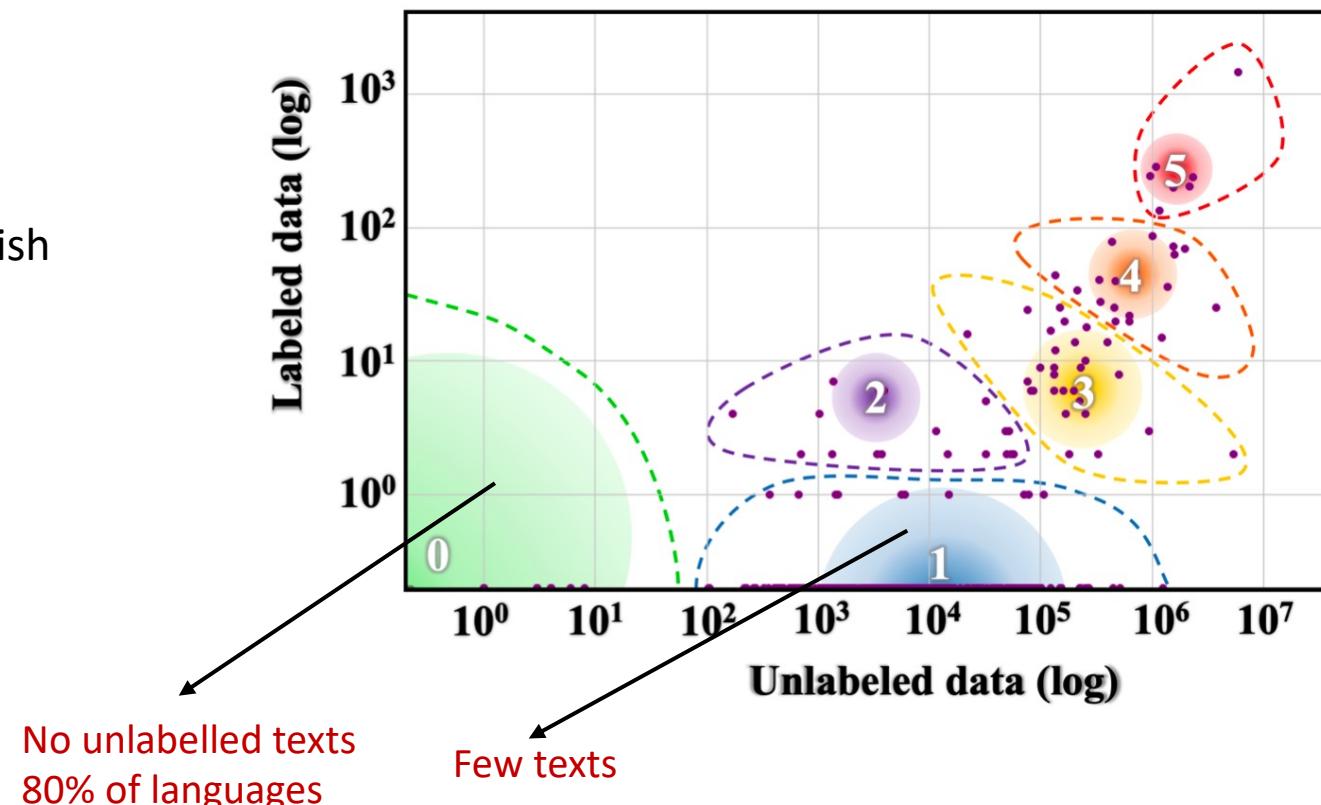


# Under-resourced languages: Labelled+Unlabelled data

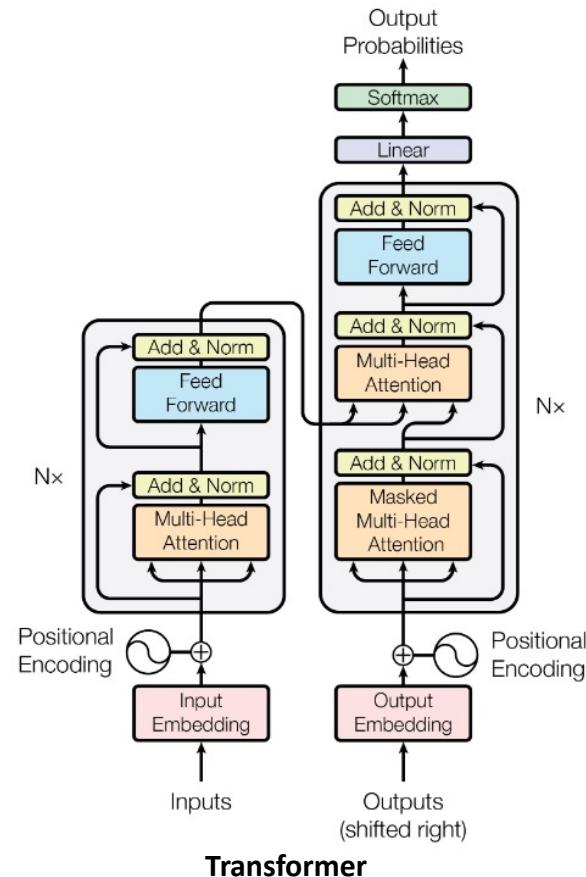
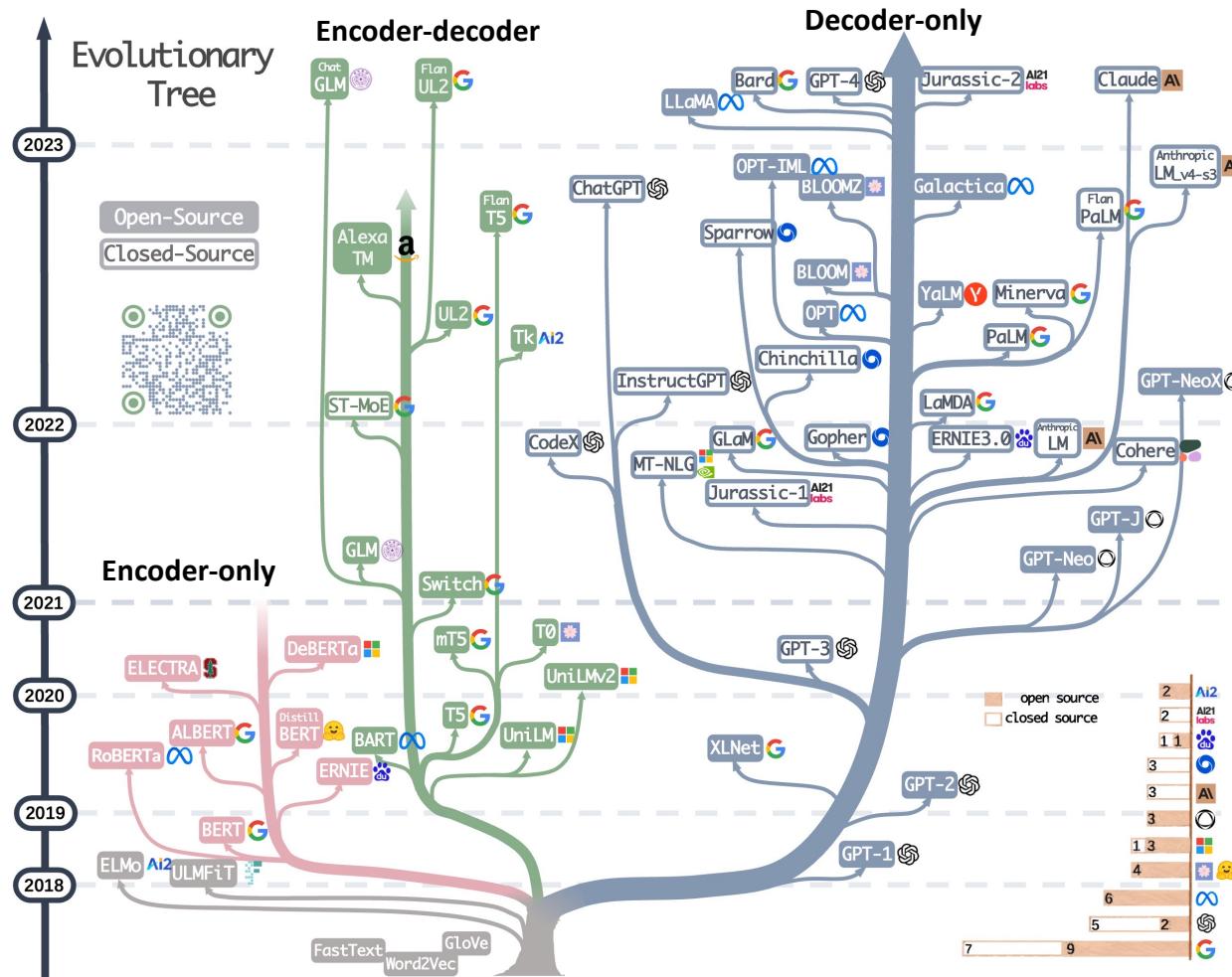
- Joshi et al (2020) proposed a **six-class categorization** of languages based on
  - Unlabelled corpora
  - Labelled corpora

**Class 5:** Winners e.g. English, Spanish

**Class 0:** Left Behind



# Multilingual Language Models



Yang et al. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. In arXiv

# NLP in a Nutshell (2013 – Today)

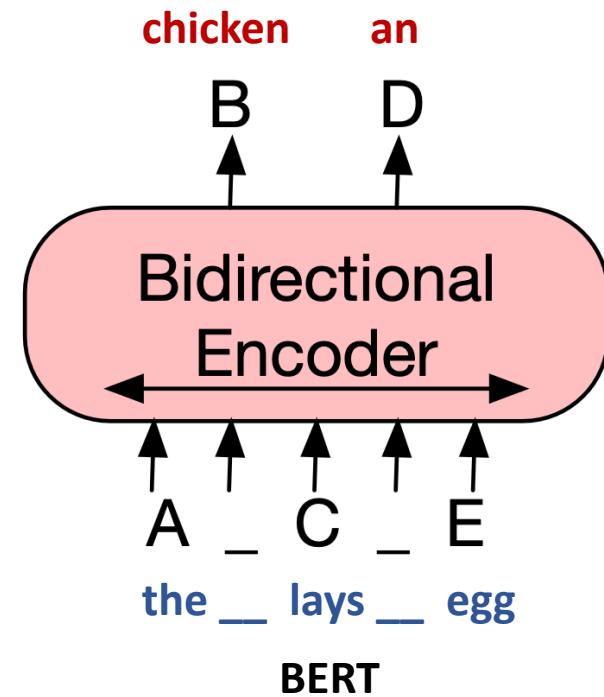
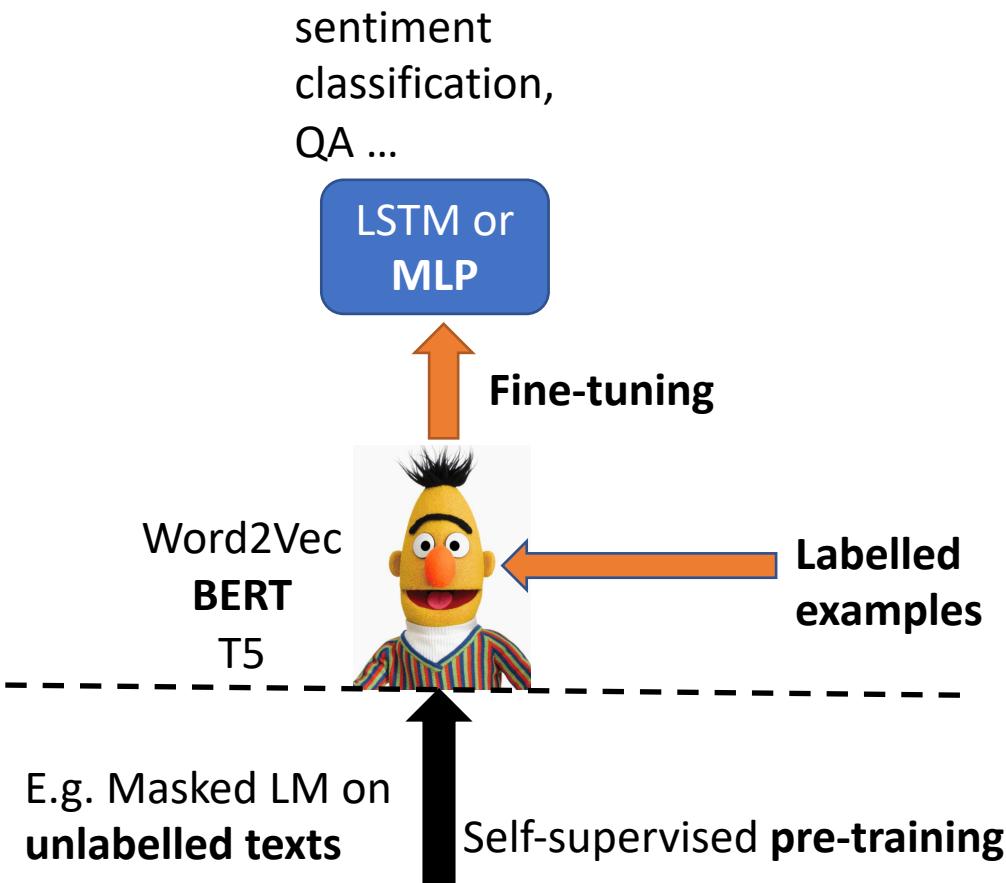


Image obtained from BART Sequence-to-Sequence Pre-training paper <https://arxiv.org/abs/1910.13461>

# NLP in a Nutshell (2013 – Today)

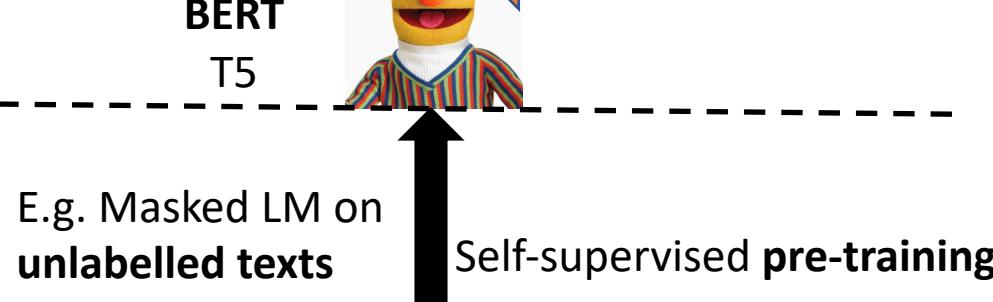
Pre-train -> **Fine-tune**

sentiment  
classification,  
QA ...

LSTM or  
MLP

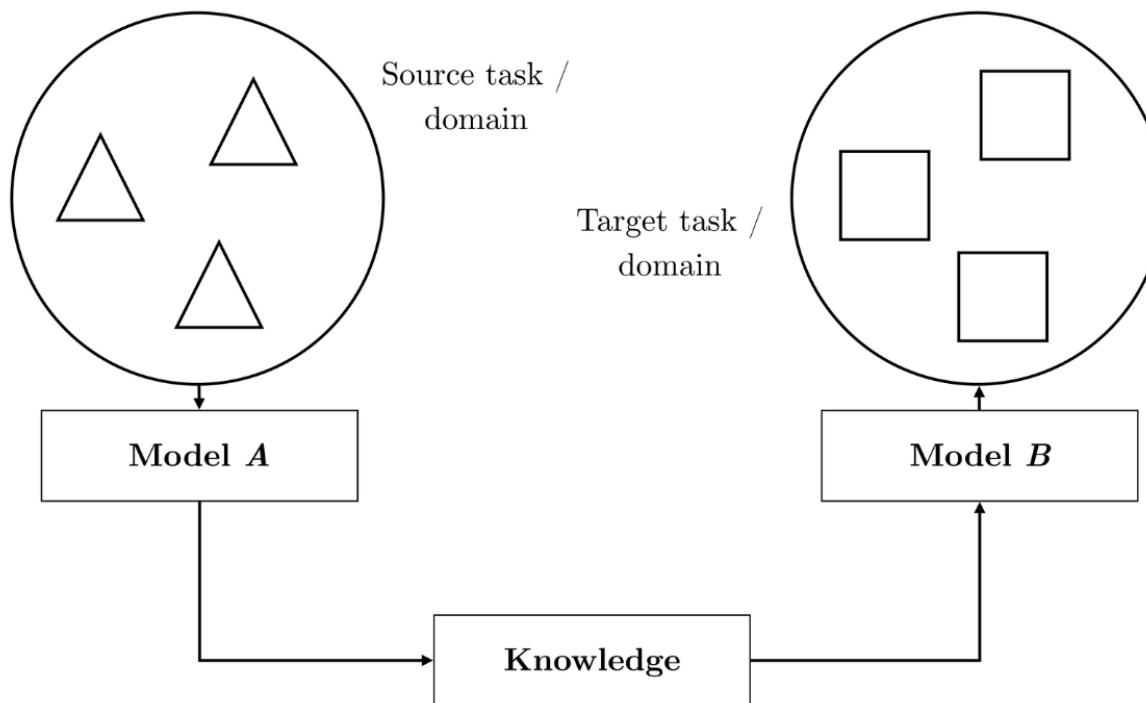
Fine-tuning

Word2Vec  
**BERT**  
T5



- **Lack of labelled data**
  - For downstream tasks
- **No large enough unlabelled data**
  - for self-supervised pre-training
  - We can only leverage **transfer learning**.

# What is transfer learning?



- Inspired by transfer learning of **ImageNet** to **other computer vision tasks**
- **Same task: labelled data in a source domain transferred to target domains.**
  - Same for **source language to target language transfer**
- **different tasks: labelled data in the target domains.**
  - Sequential learning like BERT -> sentiment classification
  - Multi-task learning

# Monolingual transfer learning

Pre-train -> **Fine-tune**

sentiment  
classification,  
QA ...

LSTM or  
MLP

Fine-tuning

Word2Vec

BERT

T5

E.g. Masked LM on  
unlabelled texts

Self-supervised pre-training

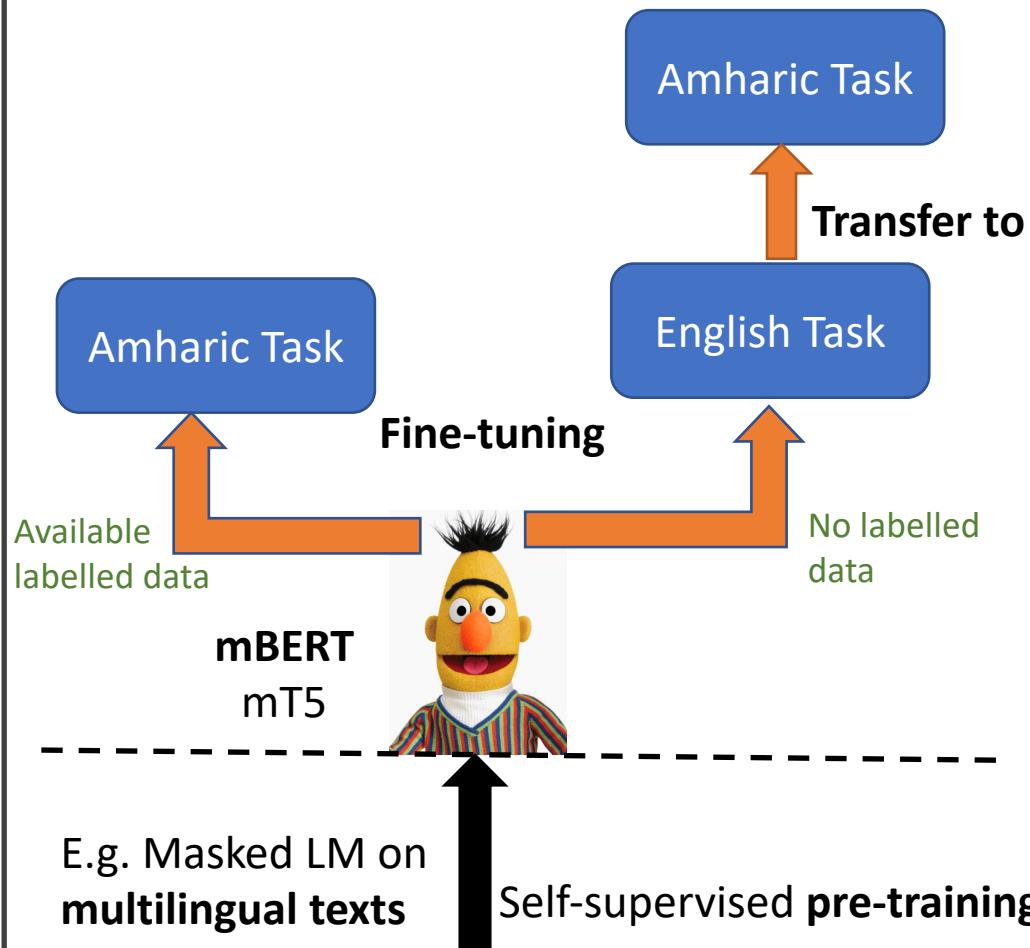
- **Feature extraction**

- Extract sentence embeddings for LSTM / CNN classifier
- Weights are frozen

- **Fine-tuning**

- End-to-end fine-tuning
- All weights are modified
- Often leads to better performance

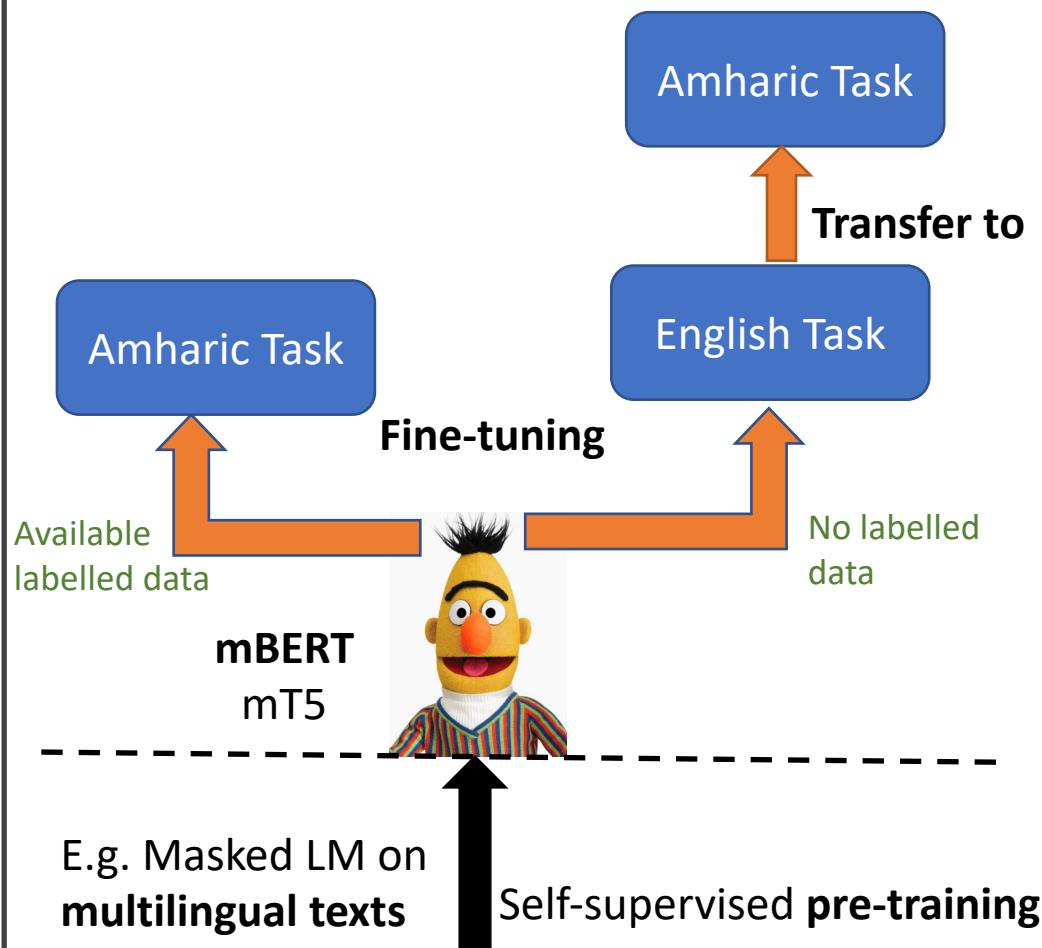
# Cross-lingual Transfer learning



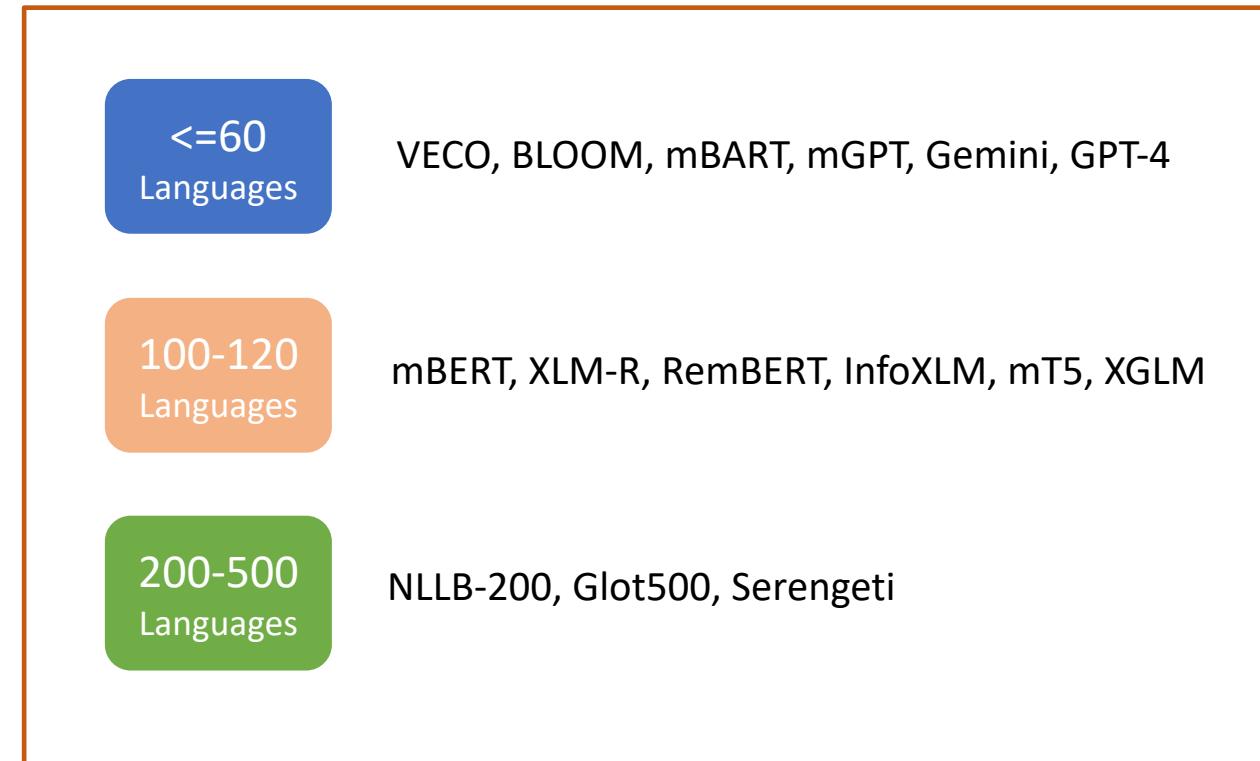
We leverage a multilingual PLM instead of a monolingual one

Suppose we want to apply **transfer learning to a new language** not covered in the pre-training of mBERT e.g. **Amharic**

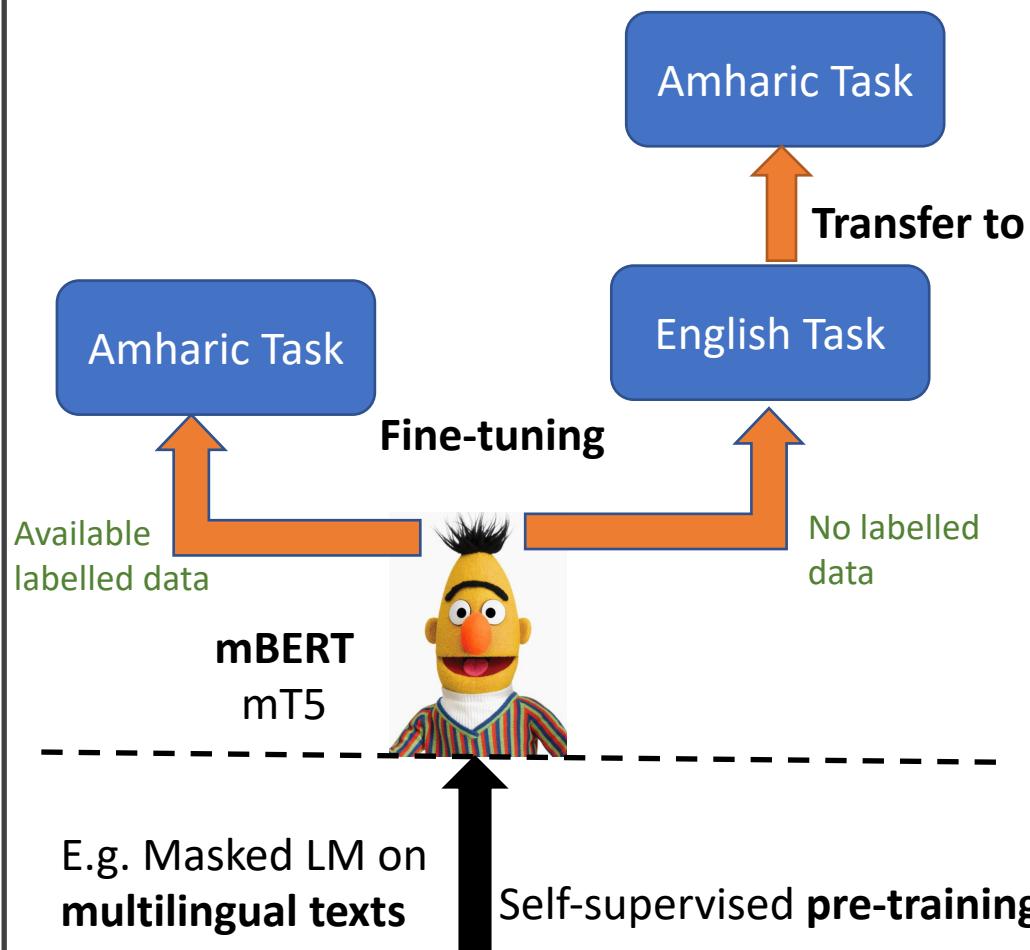
# Cross-lingual Transfer learning



Most pre-trained LMs only **cover 100 languages**, and many are not even evaluated on downstream tasks



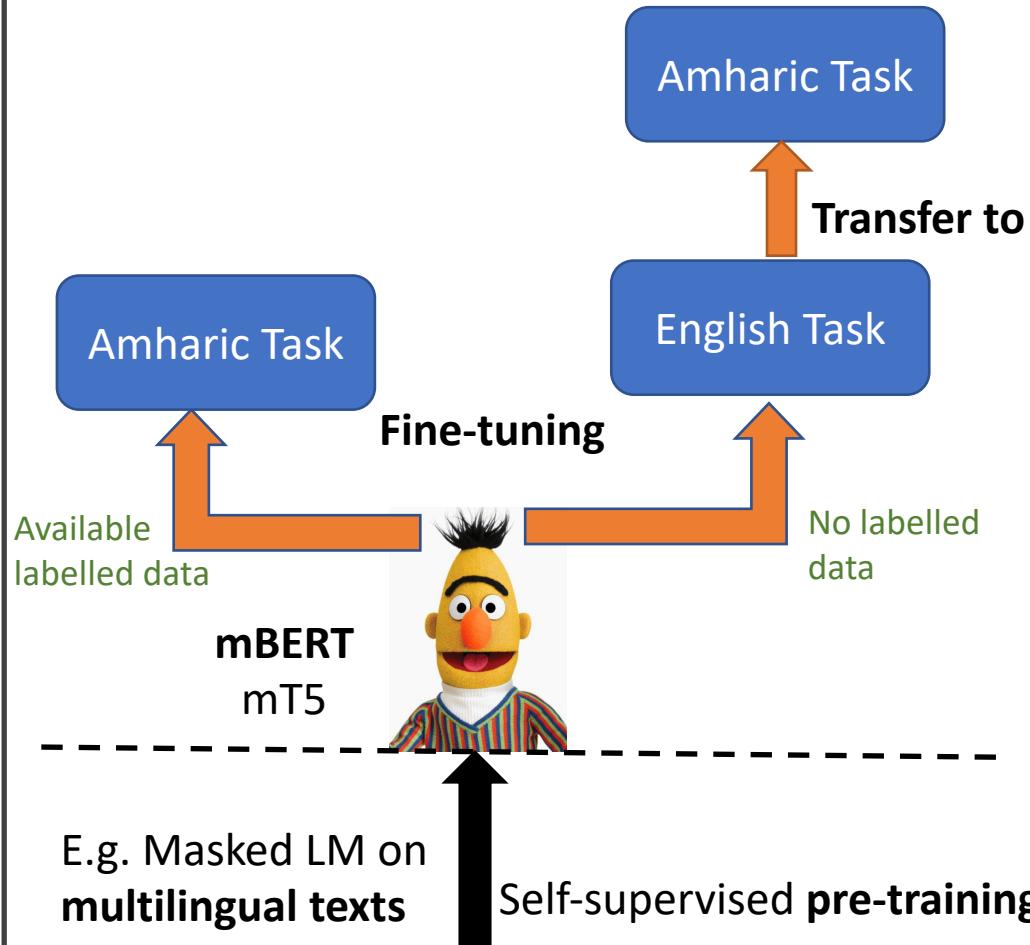
# Cross-lingual Transfer learning



Cross-lingual transfer learning is attractive because there are over **7,000+ languages**, many do **lack labelled data**



# Cross-lingual Transfer learning



## Challenges:

- Lack of **labelled datasets**
- Most **PLMs** only **cover few languages** (up to. 100)
- Source languages and target languages may use **different scripts**.
- Multiple fine-tuning leads to **catastrophic forgetting** and **parameter inefficient** (multiple checkpoints per fine-tuning runs)
- What is the **best source language** to transfer from?

# PART ONE

**Challenge 1:** Lack of labelled data for downstream tasks



**Approach to address it:** Developing labelled datasets using participatory approach

## Developing datasets for Low-resource languages

**MasakhaNER 1.0** (Adelani et al, **TACL 2021**)

**MasakhaNER 2.0**, (Adelani et al, **EMNLP 2022**)

**MasakhaNEWS**, (Adelani et al, **AACL-IJCNLP 2024**)

**SIB-200**, (Adelani et al, **EACL 2024**)

# Outline for Labelled dataset



**MasakhaNER**

African languages



***MasakhaNEWS***

*News Topic classification*

*African languages*



***SIB-200***

*200 languages*

# Outline for Labelled dataset



**MasakhaNER**

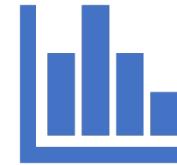
African languages



*MasakhaNEWS*

*News Topic classification*

*African languages*



*SIB-200*

*200 languages*

# Developing labelled datasets for African languages

- WikiANN
  - NER: 9 African
- CMU Wilderness
  - Text-to-Speech: > 30
- Several not free or-public
  - Linguistic Data Consortium
  - Offline

Prior to 2019

AfricaNLP Workshop (2020 - )  
Masakhane (2019 - )

- Masakha**NER**
  - 21 African languages
- Masakha**POS**
  - Part of Speech: 20 languages
- Masakha**NEWS**
  - News topic: 15 languages
- **MAFAND**
  - Machine Translation: 21 languages
- **NaijaSenti & AfriSenti**
  - Sentiment analysis: 14 languages
- **BibleTTS**
  - Text-to-Speech: 6 languages
- **AfriQA**
  - Question answering: 10 languages

In 2024

# Masakhane – Participatory Research



- **Masakhane:**
  - A grassroots NLP community **for Africans, and by Africans**
  - Pioneering Participatory research for African NLP.
  - Organizes yearly **AfricaNLP workshop** @ AI conferences.
  - Over 1900 members in slack (over **100 active**)
- Developed the first large **named entity recognition** dataset
  - 10 African languages (2020-2021)
  - **MasakhaNER**

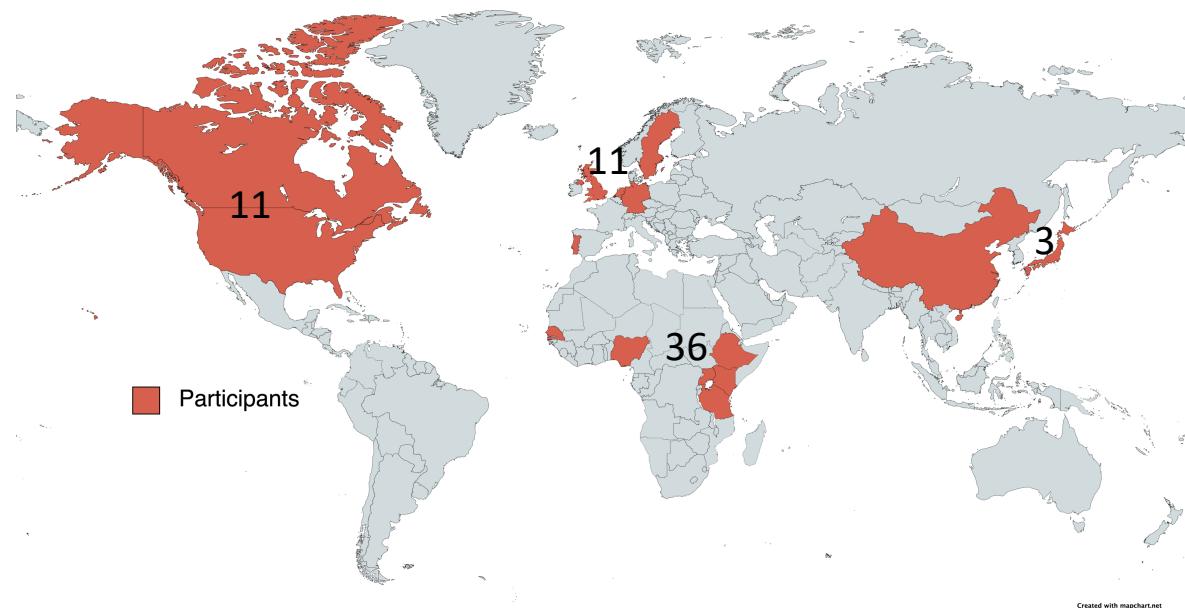
# Named Entity Recognition (NER)

- A core task in information extraction
- Recognizing entities like
  - Person,
  - Location,
  - Organization or
  - Date.
- Token classification task
- Evaluation is performed using F1-score

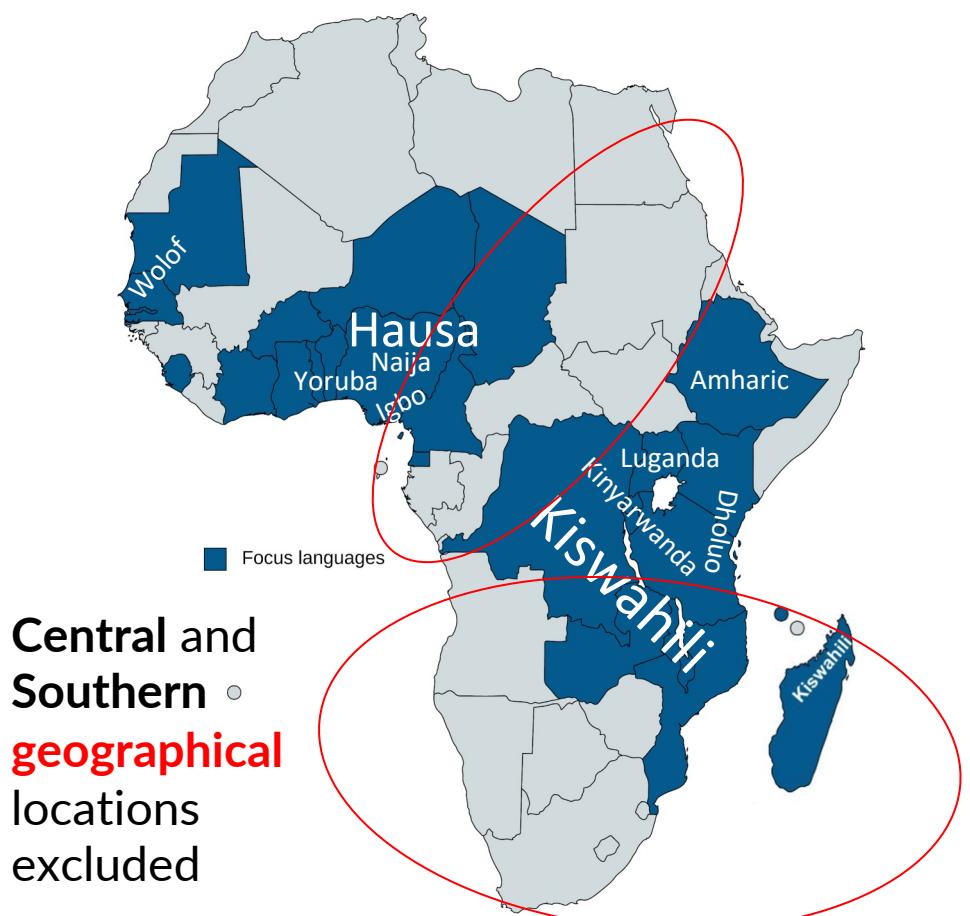
On	O
the	O
4th	B-DATE
of	I-DATE
February	I-DATE
,	O
Global	B-ORG
Voices	I-ORG
visited	O
Fernando	B-PER
Gomes	I-PER

# The MasakhaNER Project

- Comprises of different stakeholders:
  - *Native speakers*
  - *Linguists and Language experts*
  - NLP / ML experts/researchers

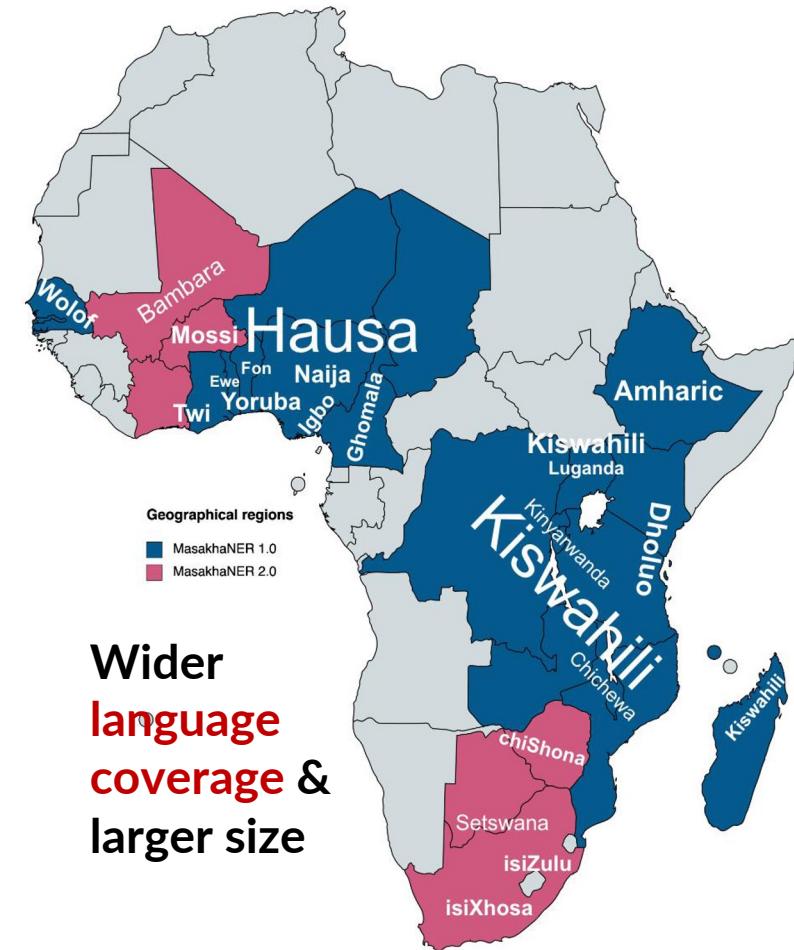


# MasakhaNER Project: two phases



**MasakhaNER 1.0:** 10 African languages

**Size:** 920 to 3,200 sentences



**MasakhaNER 2.0:** 20 African languages

**Size:** 4,800 to 11,000 sentences

# Language Characteristics in NER

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ አ.ምር በኋይደርስ ፕሮ ዓመት የአለፈውን ሁንጻን የኋ መረ አደረገት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Nigeria sarauta
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Éwé	Kano fe Emir na wobla ta na Zhang si nō Nigeria fe 18 sɔŋ la
isiXhosa	I-Emir yaseKano ithwese u-Zhang othe wachitha iminyaka eyi-18 eNigeria ngonkontsho.
isiZulu	U-Emir waseKano ubeke isigqoko kuZhang osechithe iminyaka engu-18 eNigeria .
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Émià ilú Kánò wé lágúní lé orí Zhang éni tí ó ti lo ọdún méjìdínlógún ní orílè-èdè Nàjírìà

Example of named entities in different languages. Examples includes PER, LOC, and DATE entities

Non-Latin script

Rich morphology

Rich morphology + noun classes

Extensive use of diacritics

# Outline for Labelled dataset



**MasakhaNER**

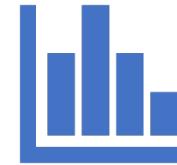
African languages



***MasakhaNEWS***

*News Topic classification*

*African languages*



***SIB-200***

*200 languages*

# News Topic Classification

- A sequence classification task
- Evaluation is performed using Accuracy /F1-score

Five tins to know about Nigeria new naira notes design

business

John Boyega don win GQ Icon Award for di second time

entertainment

AstraZeneca vaccine side effects blood clot: COVID-19 vaccine

health

UEFA apologise to Liverpool and Real Madrid fans

sports

# Geographical & Linguistic diversity of MasakhaNEWS

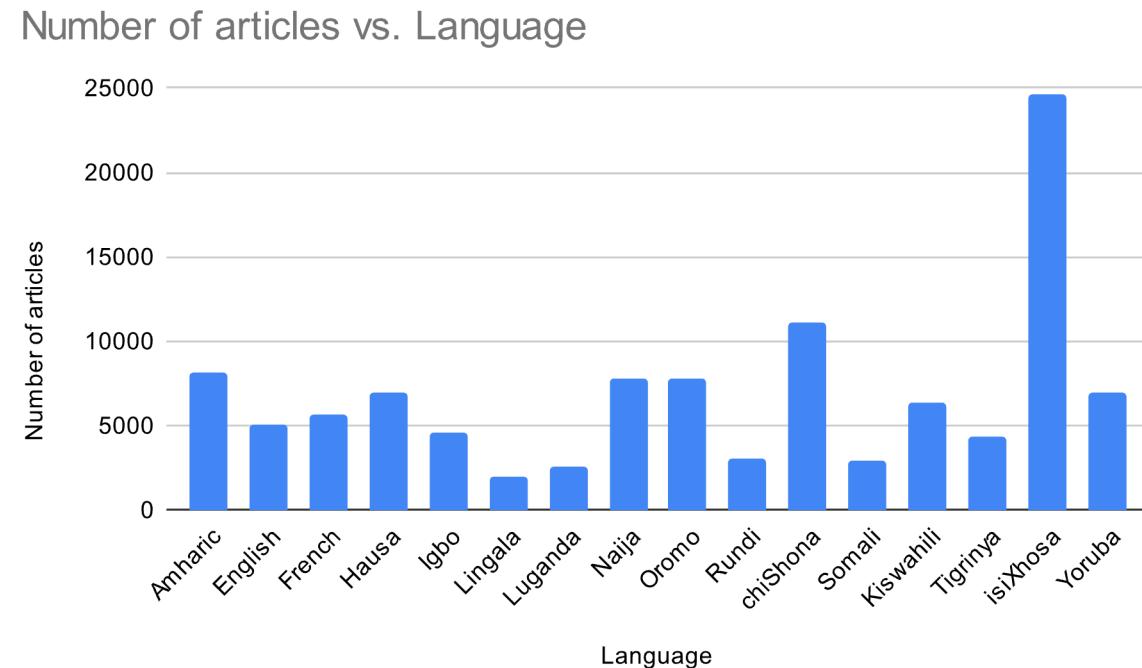


**MasakhaNEWS:** 14 African languages,  
English & French

Language	Family	Population
Amharic	Afro-Asiatic	57M
Hausa	Afro-Asiatic	77M
Igbo	Niger-Congo	31M
Lingala	Niger-Congo	40M
Luganda	Niger-Congo	11M
Naija	English Creole	121M
Oromo	Afro-Asiatic	37M
Rundi	Niger-Congo	11M
chiShona	Niger-Congo	11M
Somali	Afro-Asiatic	22M
Kiswahili	Niger-Congo	71M – 106M
Tigrinya	Afro-Asiatic	9M
isiXhosa	Niger-Congo	19M
Yoruba	Niger-Congo	46M

# MasakhaNEWS: Data Source

- BBC: 12 languages
  - English – only 5000 articles crawled
- VOA: 2 languages
  - Lingala
  - chiShona
- Other news papers:
  - Isolezwe - isiXhosa
  - Gambuuze - Luganda
  - Kwayedza – chiShona (sports)

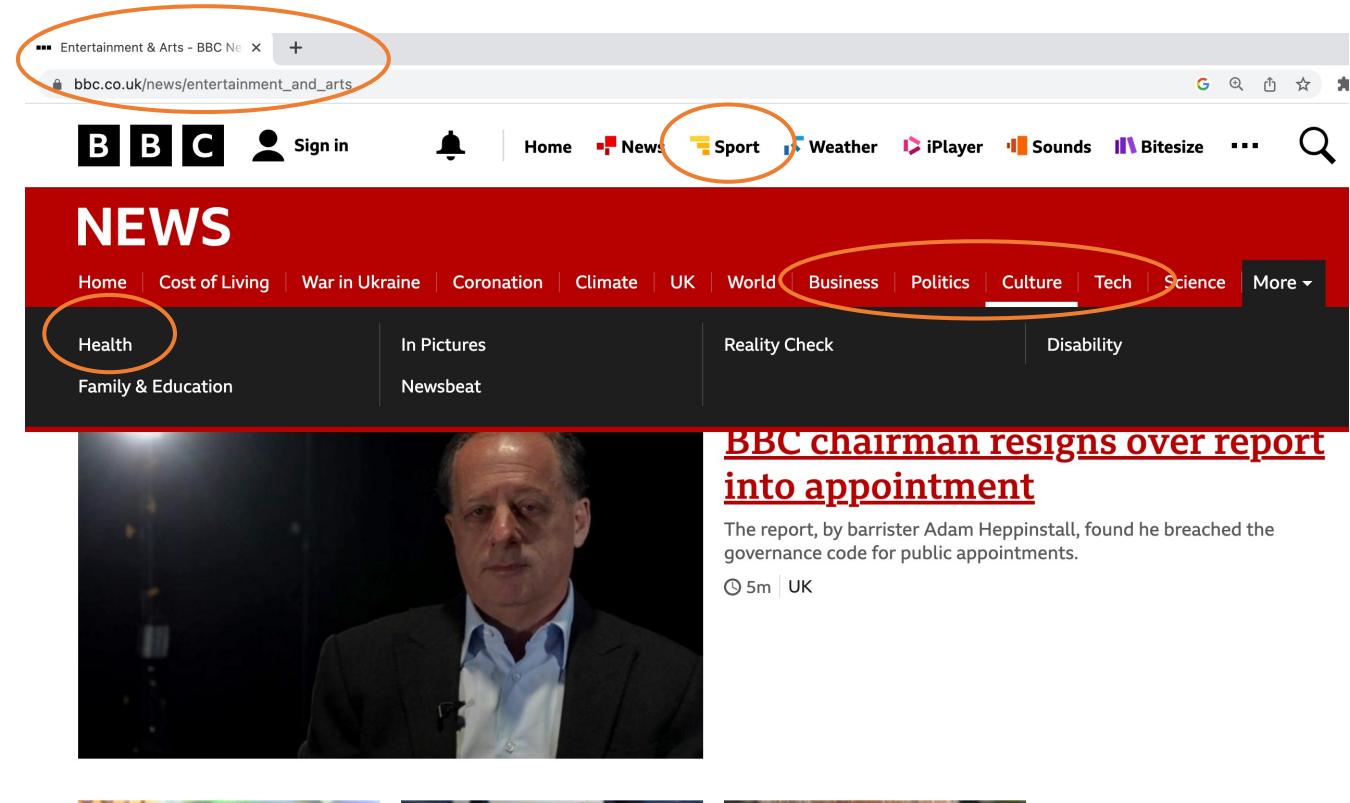


# MasakhaNEWS: Data Annotation

- **News Topics annotated:**

- Business
- Entertainment
- Health
- Politics
- Sports
- Technology
- Religion (added for prevalence in African news)

Inspired by BBC English categorization



- **Target number of sentences: 3,000**

# MasakhaNEWS: Final dataset

- **Minimum number of topics: 4**
  - At least **Health, Politics, and Sports**
  - Achieved for all languages except for **Lingala** and **chiShona**.
    - ✓ Resolved annotation conflict for Lingala
    - ✓ Collected sports news for chiShona
- **Data size:**
  - 870 – 4729 articles per language.

Language	Train/Dev/Test	# topics
Amharic (amh)	1311/ 188/ 376	4
English (eng)	3309/ 472/ 948	6
French (fra)	1476/ 211/ 422	5
Hausa (hau)	2219/ 317/ 637	7
Igbo (ibo)	1356/ 194/ 390	6
Lingala (lin)	608/ 87/ 175	4
Luganda (lug)	771/ 110/ 223	5
Oromo (orm)	1015/ 145/ 292	4
Naija (pcm)	1060/ 152/ 305	5
Rundi (run)	1117/ 159/ 322	6
chiShona (sna)	1288/ 185/ 369	4
Somali (som)	1021/ 148/ 294	7
Kiswahili (swa)	1658/ 237/ 476	7
Tigrinya (tir)	947/ 137/ 272	6
isiXhosa (xho)	1032/ 147/ 297	5
Yorùbá (yor)	1433/ 206/ 411	5

# Outline for Labelled dataset



**MasakhaNER**

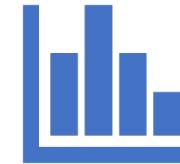
African languages



**MasakhaNEWS**

*News Topic classification*

*African languages*



**SIB-200**

**200 languages**

# Why SIB-200? Few Large-scale NLU datasets Exist

Dataset	# Languages	Task	Remark
Taxi1500 (Ma et al., 2023)	1500	Topic classification	Religious domain (Bible). Manually annotated in English, <b>parallel</b> to other languages
WikiANN (Pan et al. 2017)	176	NER	Token classification, <b>automatically annotated</b>
UD (Nivre et al., 2017)	147	POS, parsing	Manually annotated, <b>expensive annotation (time / expertise)</b> spanning several years
Belebele (Bandarkar et al. 2023)	122	Reading comprehension	Manually annotated in English, <b>parallel</b> to other languages, <b>no training data</b>
Massive (FitzGerald et al, 2023)	51	Intent / slot-filling	Manually annotated in English, <b>parallel</b> to other languages, <b>few low-resourced languages</b>

# Why SIB-200? Few Large-scale NLU datasets Exist

Dataset	# Languages	Task	Remark
Taxi1500 (Ma et al., 2023)	1500	Topic classification	Religious domain (Bible). Manually annotated in English, <b>parallel</b> to other languages
WikiANN (Pan et al. 2017)	176	NER	Token classification, <b>automatically annotated</b>
UD (Nivre et al., 2017)	147	POS, parsing	Manually annotated, <b>expensive annotation (time / expertise)</b> spanning several years
Belebele (Bandarkar et al. 2023)	122	Reading comprehension	Manually annotated in English, <b>parallel</b> to other languages, <b>no training data</b>
Massive (FitzGerald et al, 2023)	51	Intent / slot-filling	Manually annotated in English, <b>parallel</b> to other languages, <b>few low-resourced languages</b>
<b>SIB-200 (Adelani et al, 2024) (NEW)</b>	<b>205</b>	<b>Topic classification</b>	<b>General domain, Manually annotated in English, parallel to many low-resourced languages</b>

# SIB-200 dataset creation

Annotate in English and *extend annotations to other languages*

15 Labels: crime, disasters, entertainment, geography, health, nature, politics, science, sports, travel, business, religion, technology, education, uncategorized

1578	health	Check the label for specific first aid instructions for that specific poison.
996	travel	In all cases, you must book by phone directly with the airline.
197	health	The president's condition is stable, though he will be isolated at home for several days.
460	entertainment	Most televisions are made in a way to please the general public.

Arabic		
1578	health	شوف الملصق باش تتعزف على إرشادات الإسعافات الأولية الخاصة باسم المحدد.
996	travel	في جميع الظروف، يلزمك تحجز بالهاتف مباشرةً مع شركة الطيران.
197	health	ماقماش علاش بش تخافو، حالة الرئيس مستقرة، أما بش نعزلوه وحدو في الدار أيامات.
460	entertainment	تتصنع معظم التلفاز بطريقة ترضي عامة الجمهور.

Yoruba		
1578	health	Se àyéwò lébèlì náà fún ìtọhísoñà ìrànlówo' àkókó' fún májèlé ọhún ní pàtó.
996	travel	Ní gbogbo igba o gbodo gbàyè sílẹ́ pèlú ilé ise ofurufu lórí ago.
197	health	Ara ààrè tì balé, lóóótó' ó sì má a wà ní iyaraenisoñó' nílē fún ọpọlọpó' ojo'.
460	entertainment	Ọpó ेró amóhùnmáwòrán ni wọn şédá láti té àwọn ará ilú lórùn.

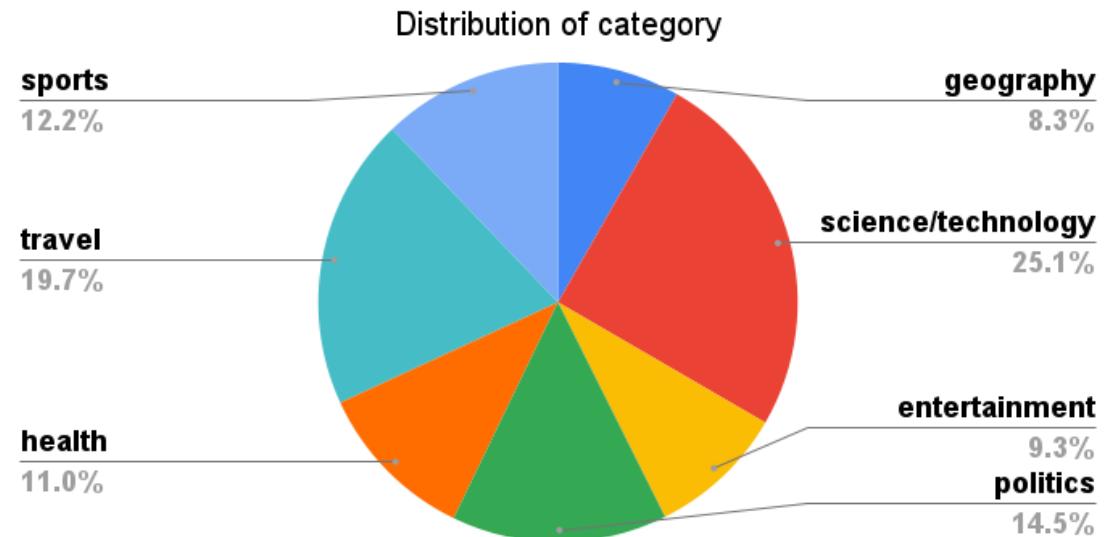
Maltese		
1578	health	Iċċekkja t-tikketta għal struzzjonijiet speċifici tal-ewwel għajnuna għal dak il-velenu speċifiku.
996	travel	Fil-każijiet kollha, trid tibbukkja mal-linja tal-ajru direttament bit-telefon.
197	health	Il-kundizzjoni tal-president hija stabbli, għalkemm huwa se jkun iżolat id-dar għal diversi jiem.
460	entertainment	Ħafna televiżjonijiet isiru b'tali mod li l-pubbliku ġenerali jogħġebu.

# SIB-200 annotation

Annotated by four native English speakers over 2 weeks

Fleiss Kappa score: 0.44 (moderate annotation agreement)

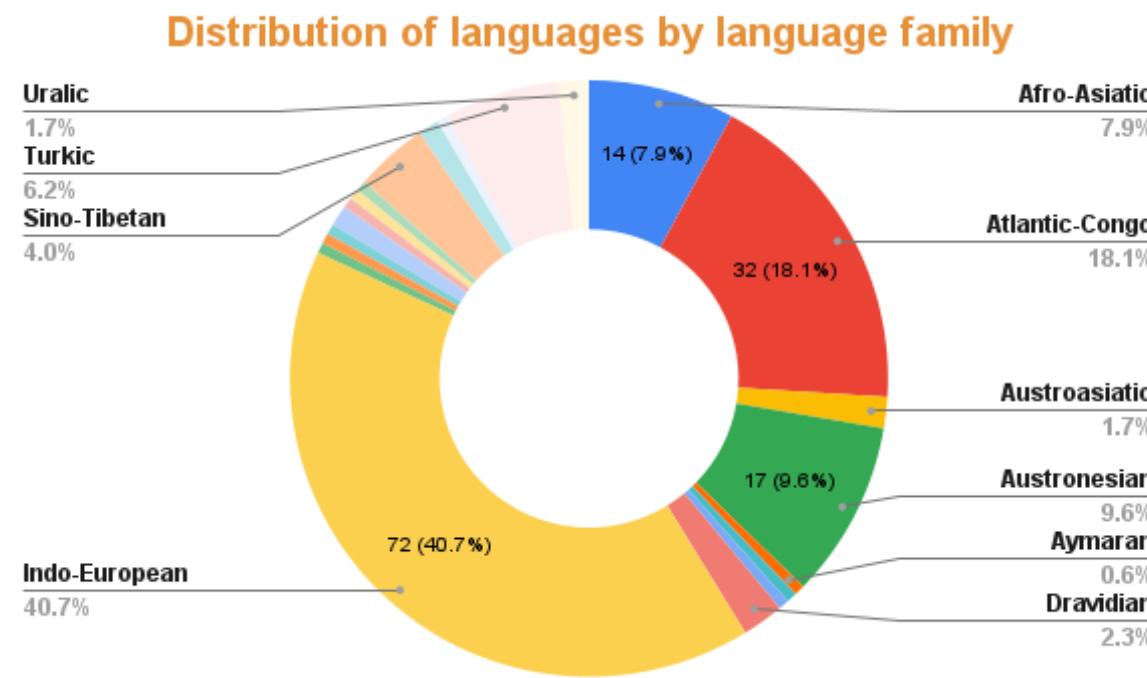
Label	TRAIN	DEV	TEST	TOTAL
science/technology	176	25	51	252
travel	138	20	40	198
politics	102	14	30	146
sports	85	12	25	122
health	77	11	22	110
entertainment	65	9	19	93
geography	58	8	17	83
Total	701	99	204	1,004



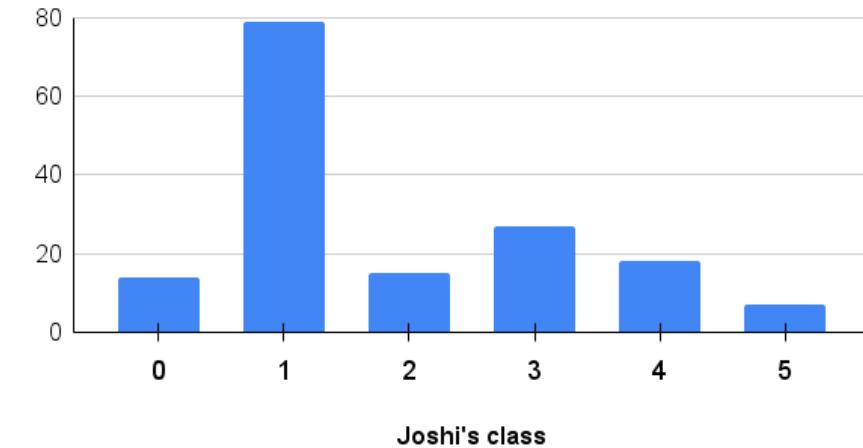
We excluded infrequent classes (<81), we also release another version of the dataset with 14 classes (excluding "uncategorized").

# SIB-200 distribution of languages

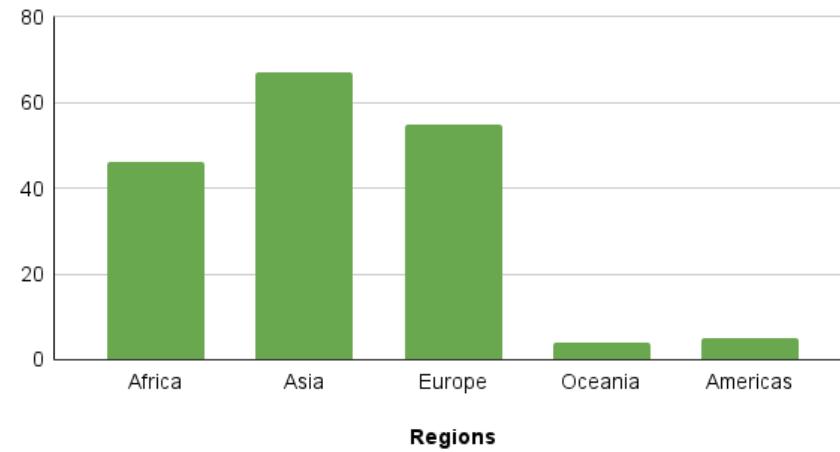
Distribution of languages covered by language family, Joshi's class and regions.



**Distribution of languages by Joshi's class**

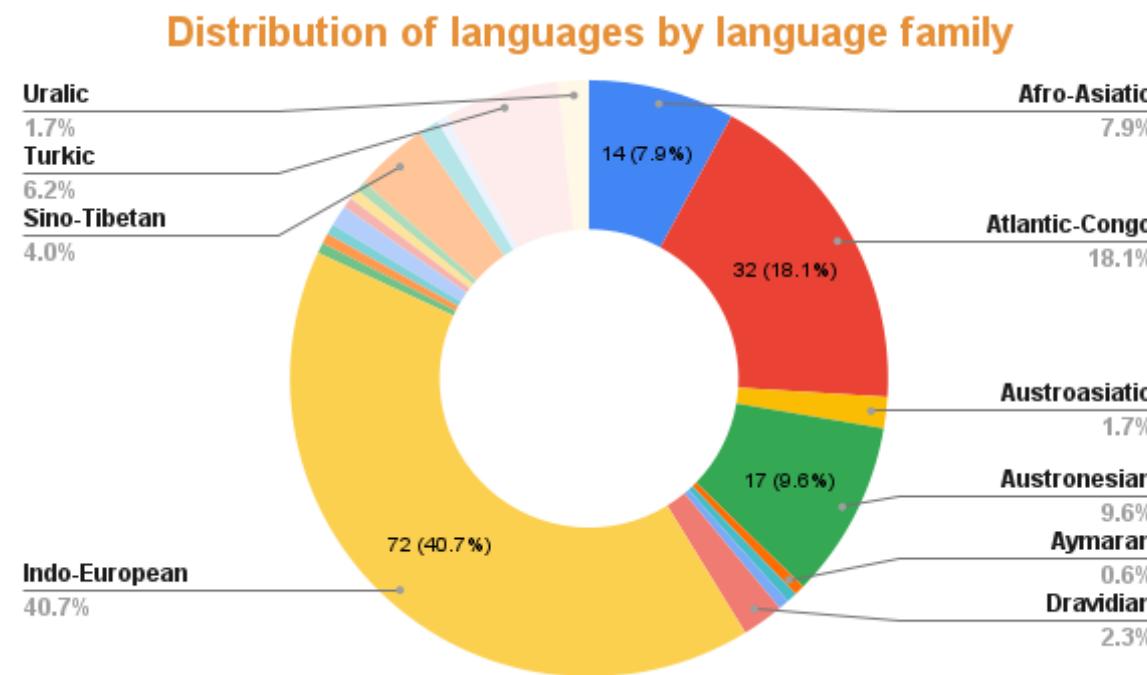


**Distribution of languages by Regions**

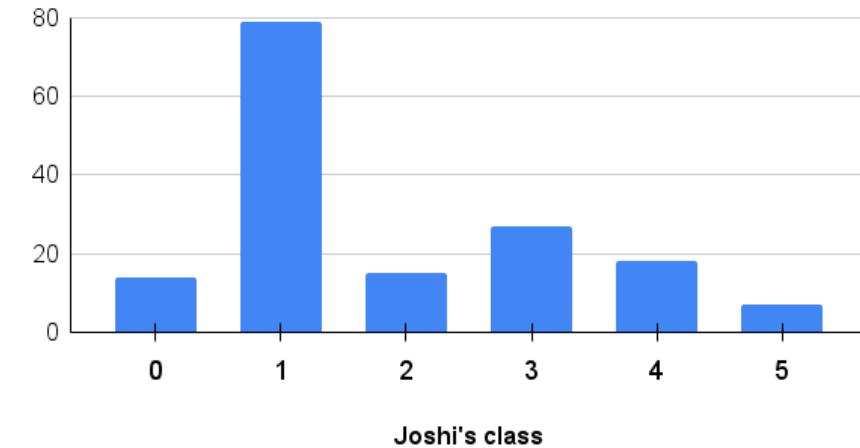


# SIB-200 distribution of languages

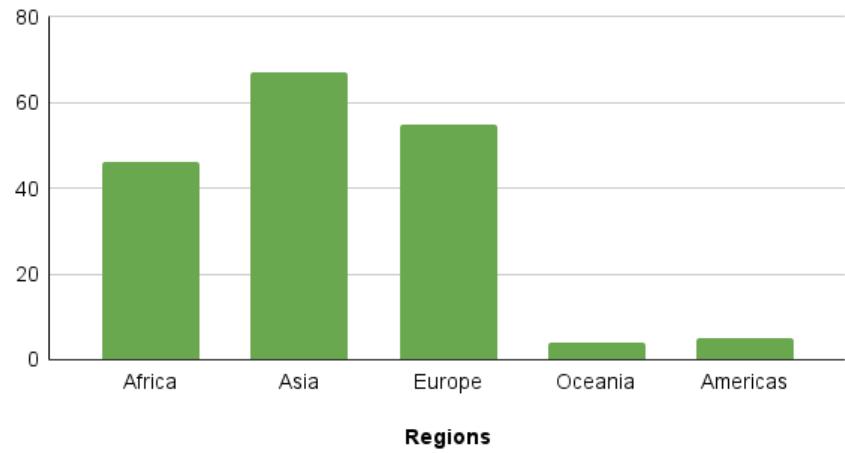
Distribution of languages covered by language family, Joshi's class and regions.



**Distribution of languages by Joshi's class**



**Distribution of languages by Regions**



# Baseline Models

Classical ML

MLP

XLM-R

Glot-500

Prompting LLMs

Pretrained language models (PLM)

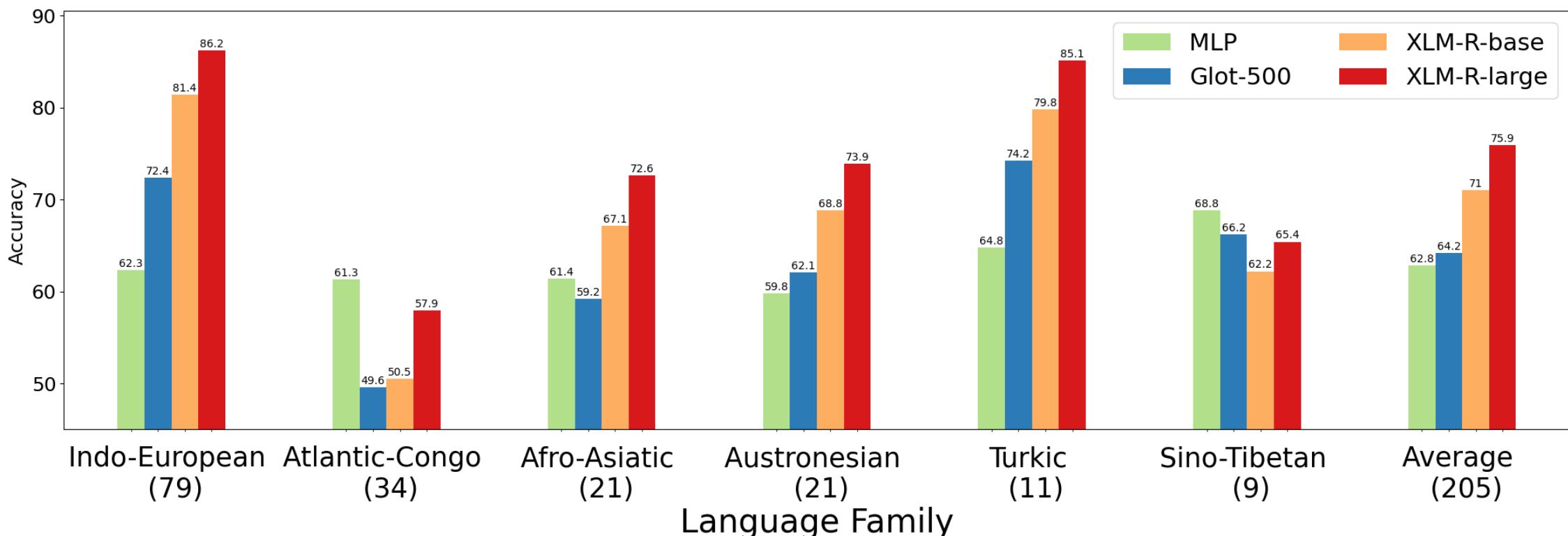
Supports 90 & 177 languages respectively.

Fine-tuned for topic classification

*Is this a piece of news regarding {"science, technology, travel, politics, sports, health, entertainment, or geography"}? {{INPUT}}*

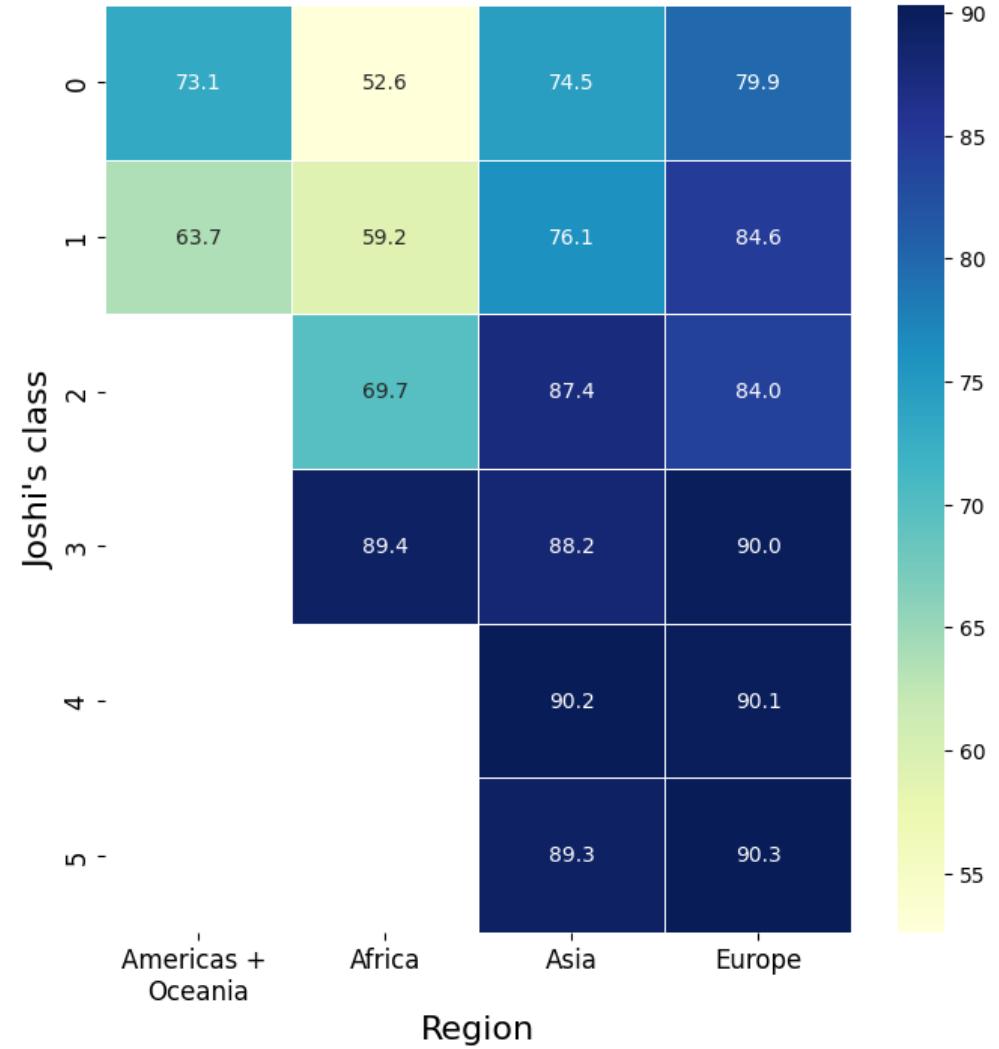
# Result: Accuracy by language family

- XLM-R provides the **best** overall result
- Simple MLP is better for 5 (out of 21) language families: Atlantic Congo, Sino-Tibetan, Nilotc, Mande, & Aymaran
  - This *language families are often not covered* in the pre-training of XLM-R



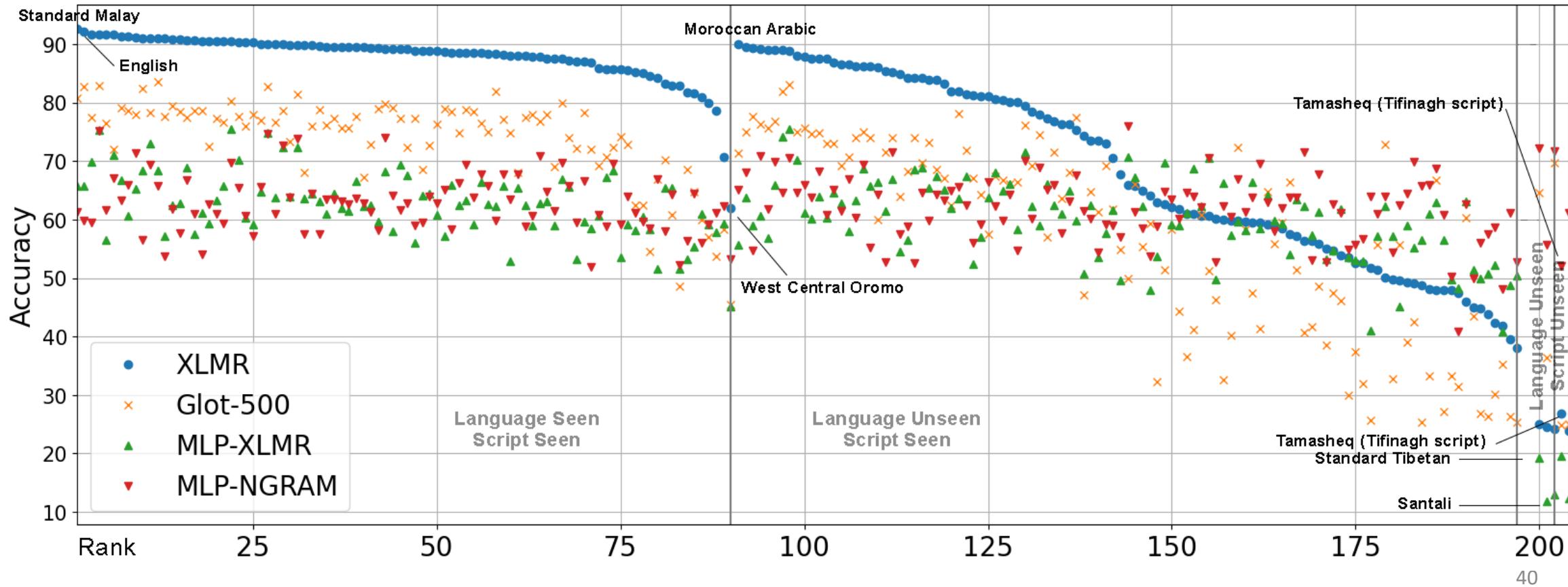
# Result: Accuracy by Joshi's class and region

- **High-resource languages:** Joshi's class 4 & class 5 have **high accuracy (>85%)**
- **Middle-resource languages:** Joshi's class 3 also have **high accuracy**
  - Due to the simplicity of the task
- **Low-resource languages:** Joshi's class 0 to 2 , **performance varies per region**
  - European and Asian languages have a **moderate accuracy**
  - **African, Oceania and Americas** regions have **poor results**



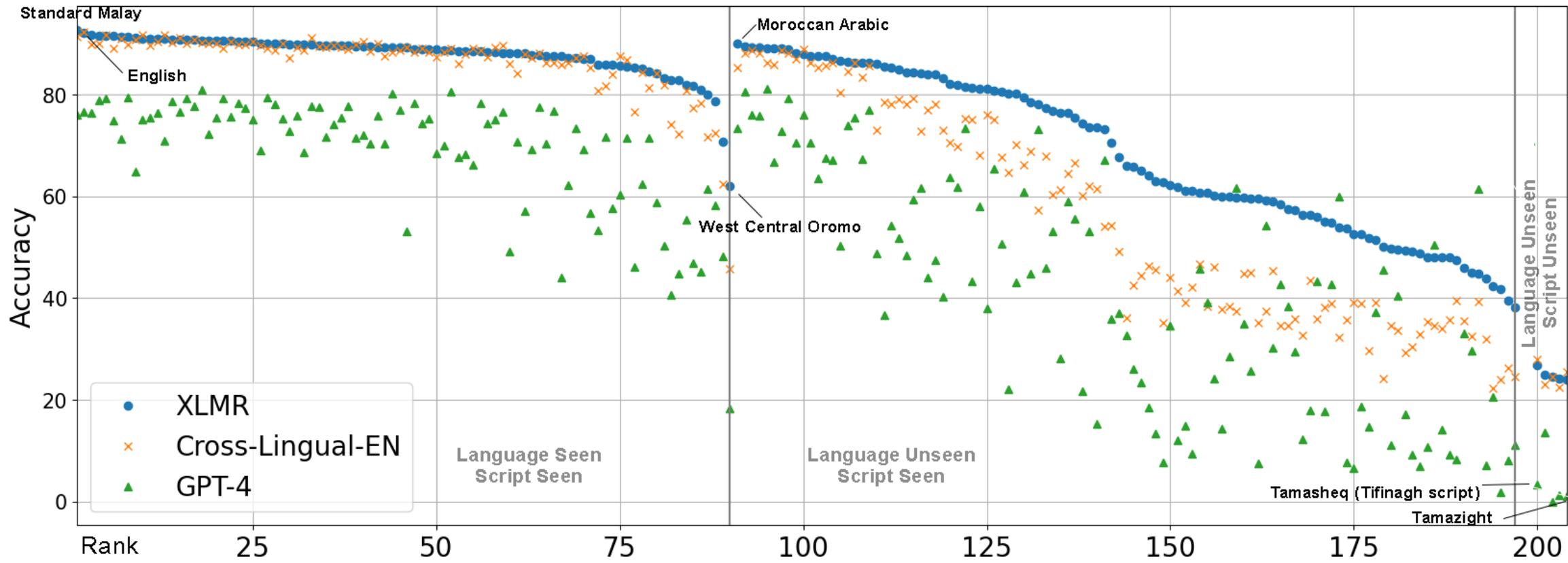
# Factors affecting performance: Language coverage

Languages unseen especially with unseen script have the worst performance (e.g. Tamazight)



# Zero-shot LLM prompting vs cross-lingual transfer

GPT-4 performance is consistently worse than cross-lingual transfer from English and in-language training



# PART TWO

**Challenge 2:** Limited coverage of PLM for new languages and scripts

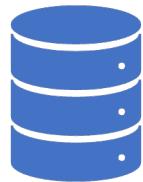


**Approach to address it:** Developing or adapting multilingual language models

## Adapting Pre-trained Language Models to a Region of Languages

AfroXLMR (Alabi & Adelani et al, COLING 2022)

# Outline for AfroXLMR



**Challenges of  
adapting PLM**

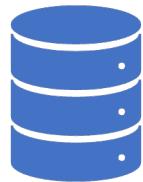


***Method***  
*Multilingual adaptive  
fine-tuning*



***Results***

# Outline for AfroXLMR



**Challenges of  
adapting PLM**



*Method*  
*Multilingual adaptive  
fine-tuning*



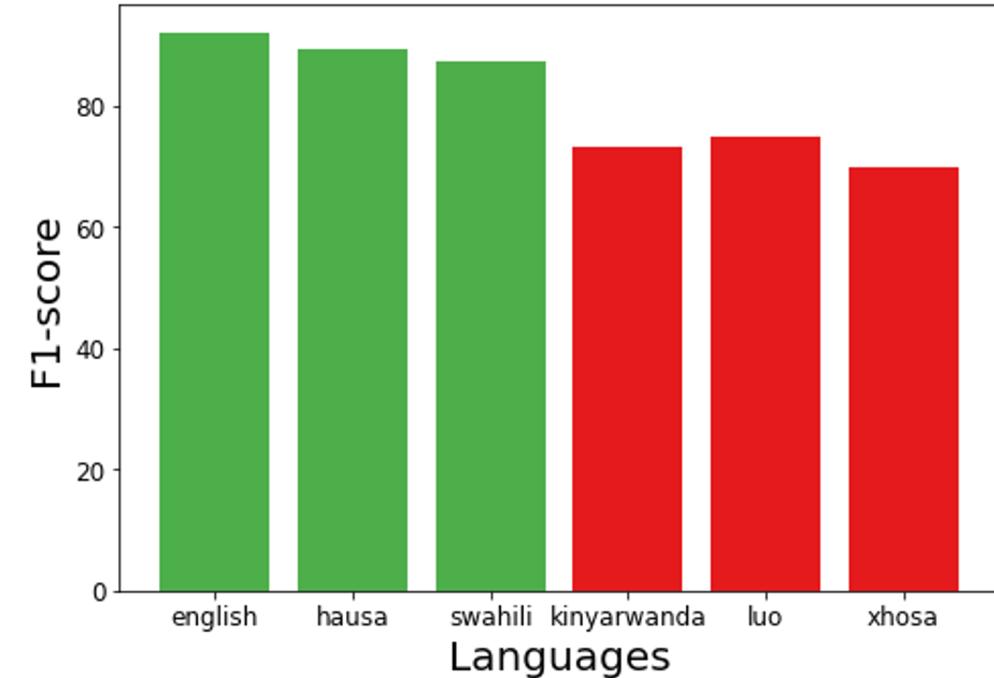
*Results*

# The Challenge of adapting PLMs to new languages

## 1. Limited language coverage in PLMs

- Curse of multilinguality
- Non-availability of corpora
  - Most PLMs limit coverage to 100 languages

→ Significant performance drop on unseen languages



Performance of PLM on MasakhaNER data  
Kinyarwanda, luo, xhosa were not used in pretraining

# The Challenge of adapting PLMs to new languages

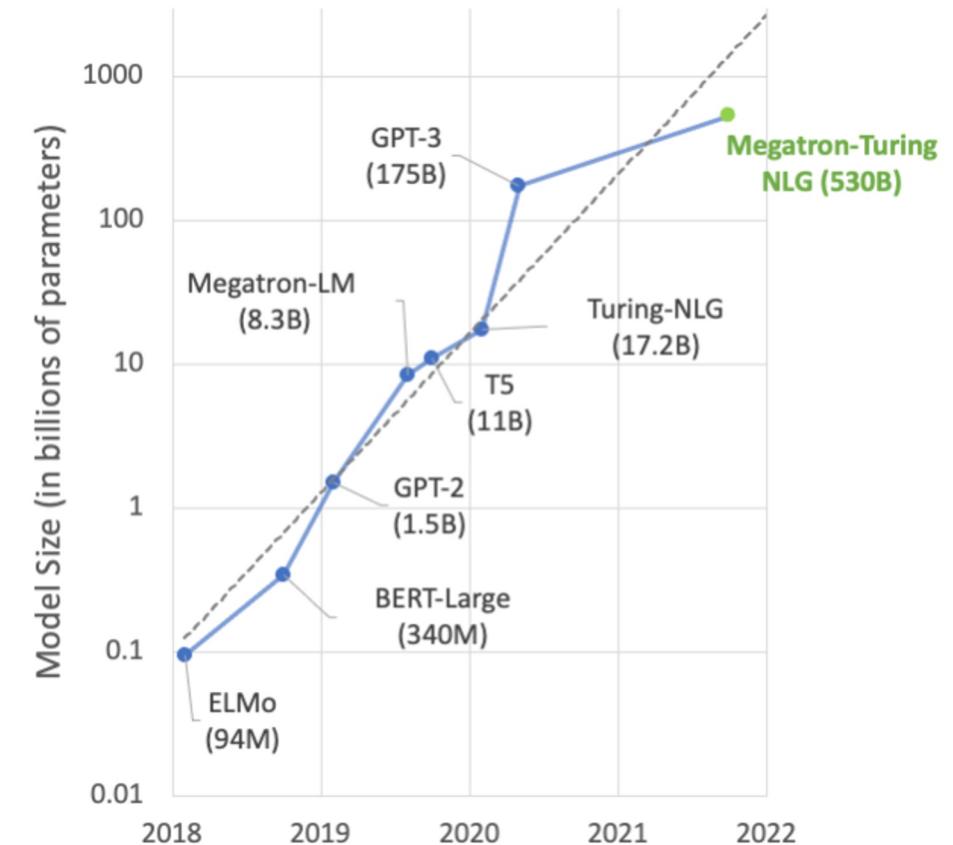
## 1. Limited language coverage in PLMs

- Curse of multilinguality
- Non-availability of corpora
  - Most PLMs limits coverage to 100 languages

→ Significant performance drop on unseen languages

## 2. Hardware restriction of PLMs

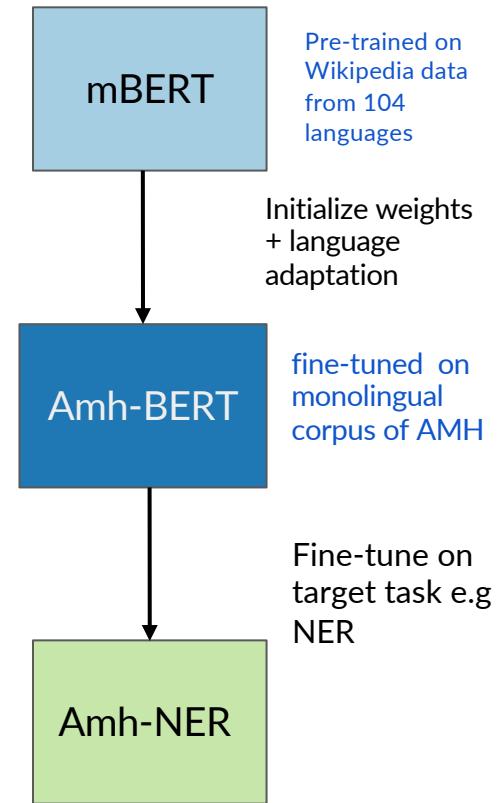
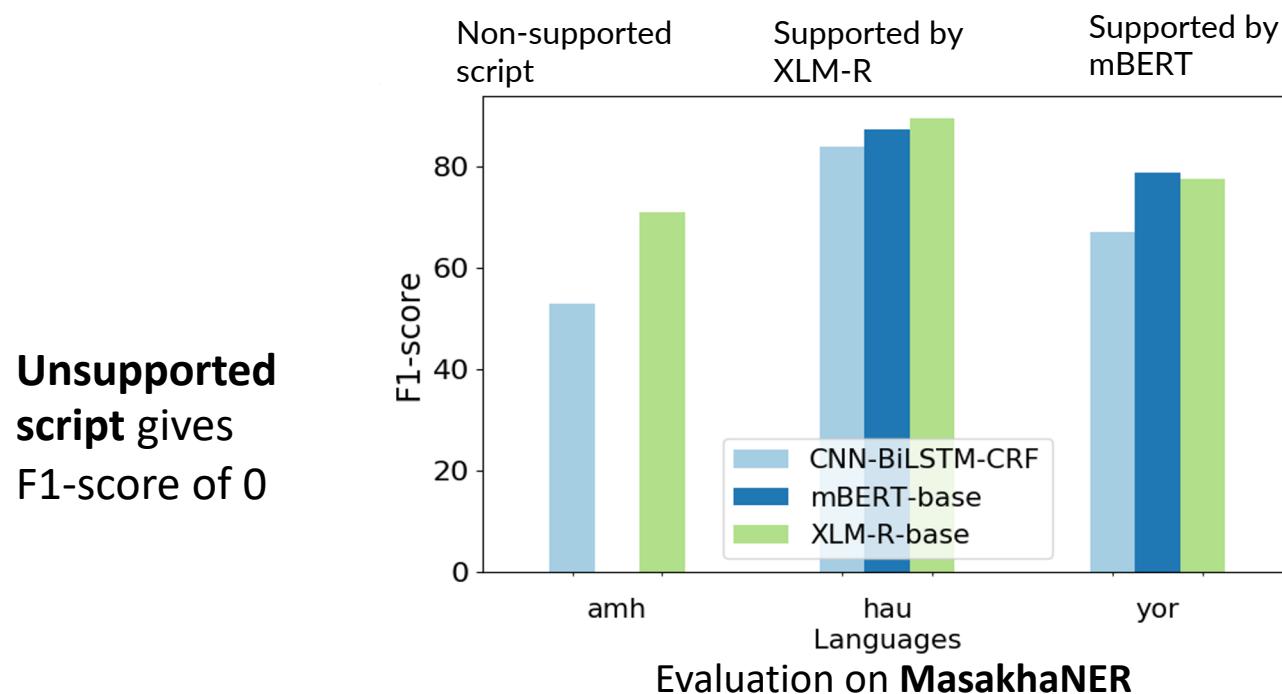
- PLMs are getting bigger and better
- Finetuning and deployment becomes a constraint



# Addressing “Limited Language Coverage” with LAFT

Perform *language adaptation* using multilingual PLM

- Language Adaptive Fine-tuning (LAFT)
- Fine-tune a PLM on monolingual corpus using the same pre-training objective

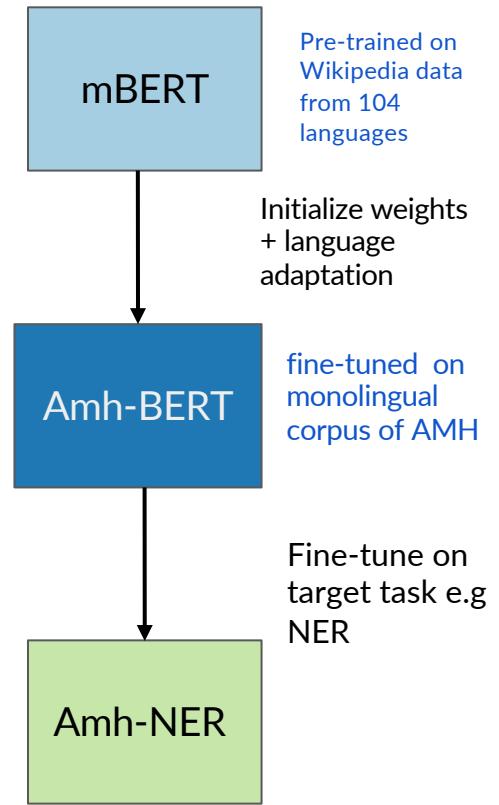
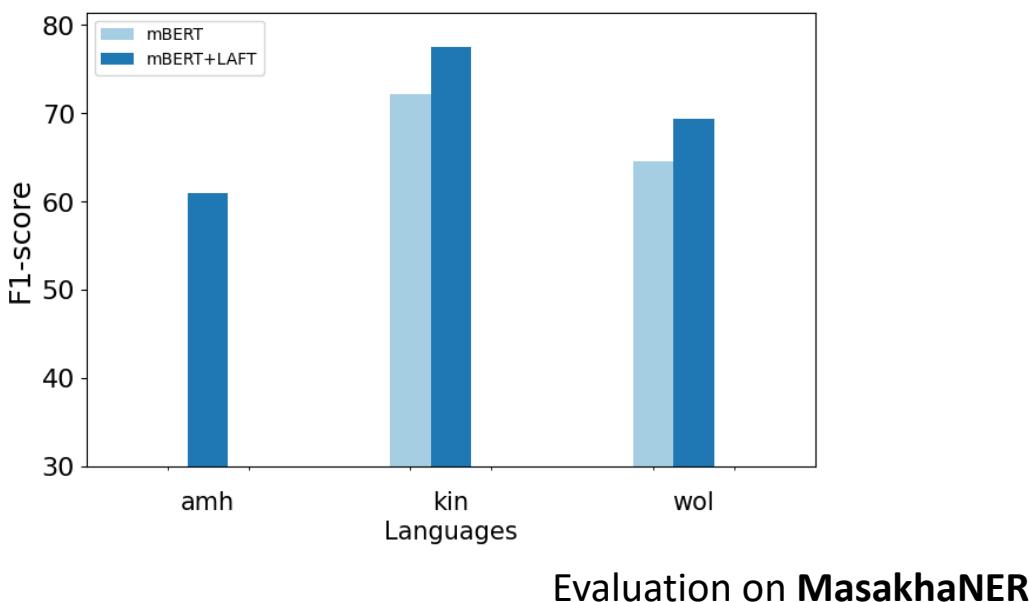


# Addressing “Limited Language Coverage” with LAFT

Perform *language adaptation* using multilingual PLM

- **Language Adaptive Fine-tuning (LAFT)**
- Fine-tune a PLM on monolingual corpus using the same pre-training objective

LAFT improves over languages

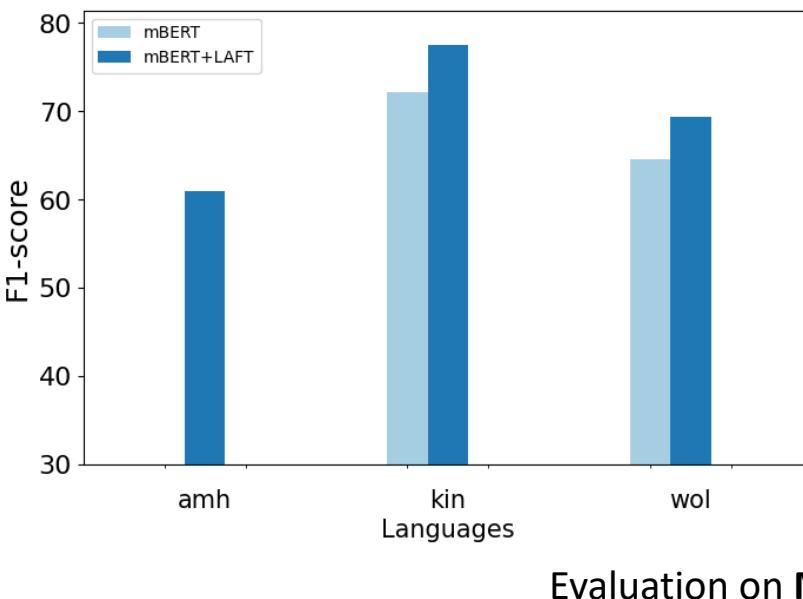


# Addressing “Limited Language Coverage” with LAFT

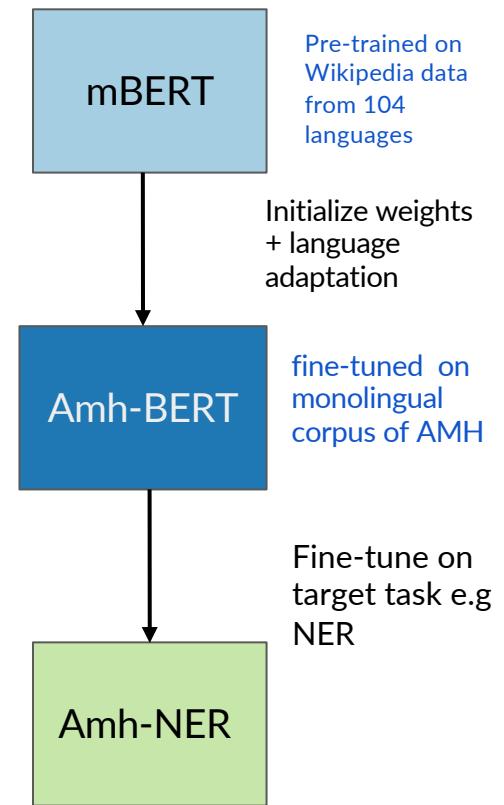
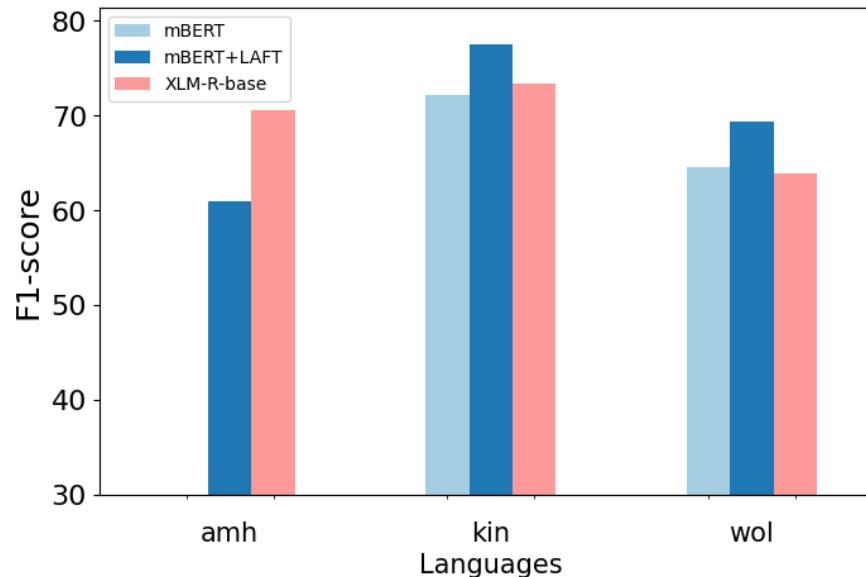
Perform *language adaptation* using multilingual PLM

- **Language Adaptive Fine-tuning (LAFT)**
- Fine-tune a PLM on monolingual corpus using the same pre-training objective

LAFT improves over languages



Models still struggle if the PLM has not been pre-trained on the script



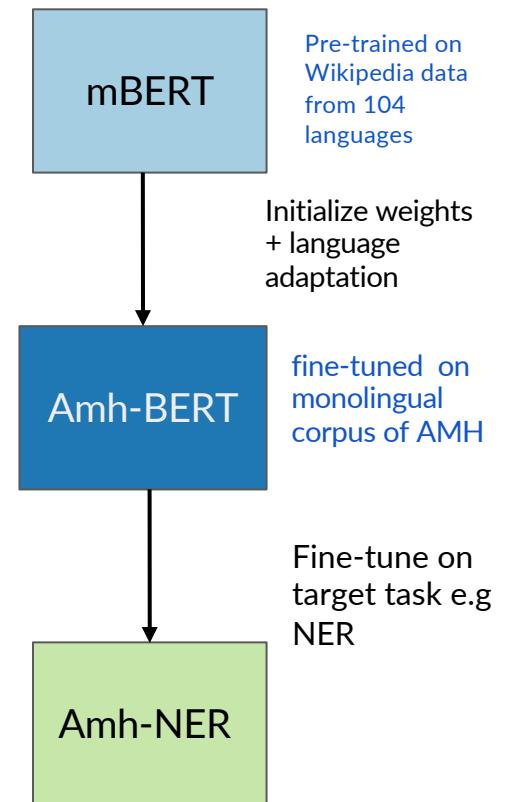
# Addressing “Limited Language Coverage” with LAFT

Perform *language adaptation* using multilingual PLM

- **Language Adaptive Fine-tuning (LAFT)**
- Fine-tune a PLM on monolingual corpus using the same pre-training objective

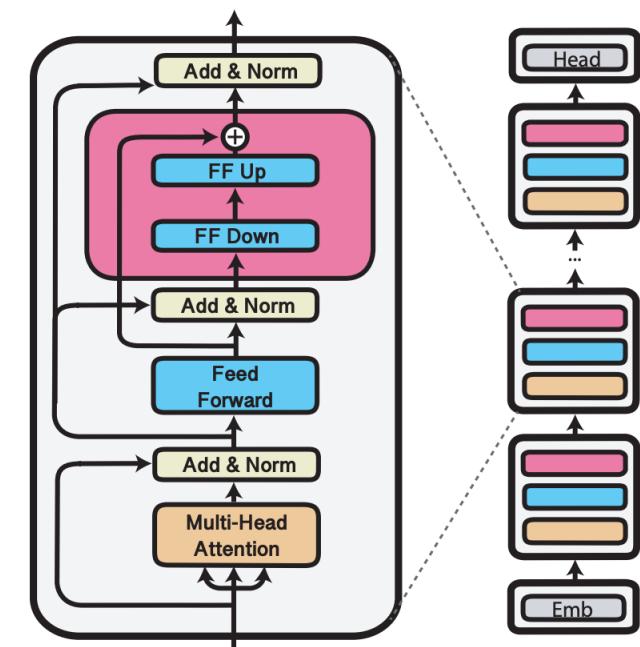
## Limitations of LAFT

- **Difficult to adapt for non-supported scripts**
- **LAFT creates a copy** of the PLM model for each language
  - **parameter inefficient** for many languages.
- **Limit the cross-lingual transfer abilities** of the adapted model



# How to address other limitations of LAFT?

1. Pre-train a multilingual model that covers **more languages** and **several scripts**
  - E.g. **AfriBERTa** (Ogueji et al., 2021) or MuRIL (Khanuja et al. 2021)
  - **Resource intensive (compute and data requirements)**
2. Use a **more parameter-efficient approach**
  - Like **Adapters** (Pfeiffer et al., 2020)
    - More parameter-efficient
    - **Still requires adaptation per language (same as LAFT)**



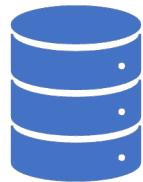
# Effectiveness of region-specific pre-training

1. Pre-train a multilingual model that covers **more languages** and **several scripts**
  - E.g. **AfriBERTa** (Ogueji et al., 2021) or MuRIL (Khanuja et al. 2021)
  - **Resource intensive (compute and data requirements)**

Models	Language Family				
	Indo-E (18)	Dravidian (4)	Austro-Asia (1)	Sino-Tib (1)	All (24)
XLM-R	86.5	87.9	24.6	48.7	82.6
IndicBERTv2	85.4	88.3	<b>65.5</b>	43.2	83.3
MurilBERT	<b>87.5</b>	<b>89.9</b>	23.5	<b>66.3</b>	<b>84.4</b>

Evaluation on SIB-200

# Outline for AfroXLMR



Challenges of  
adapting PLM



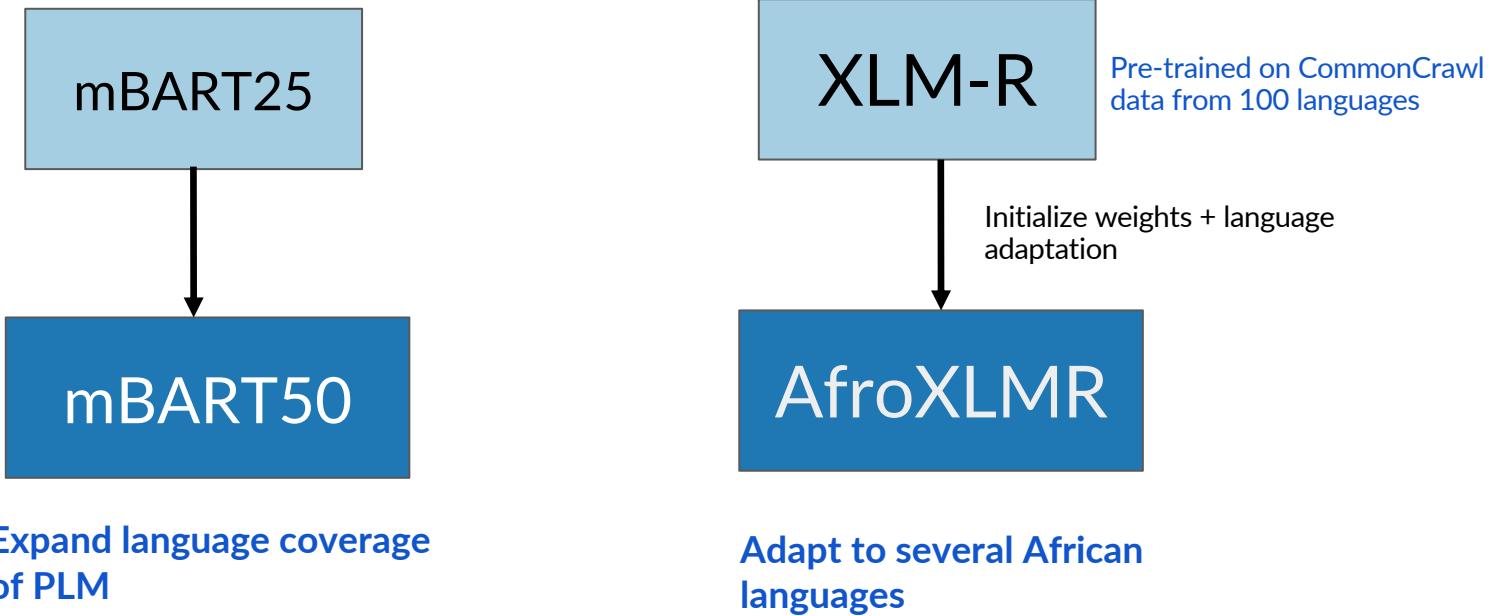
***Method***  
*Multilingual adaptive  
fine-tuning*



*Results*

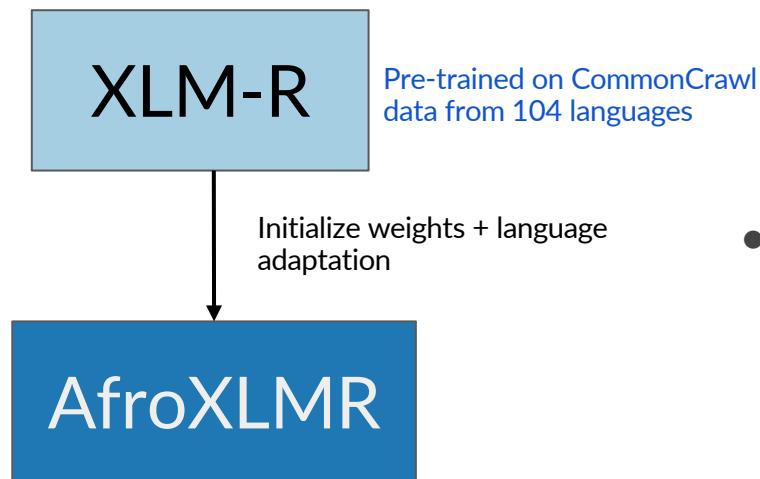
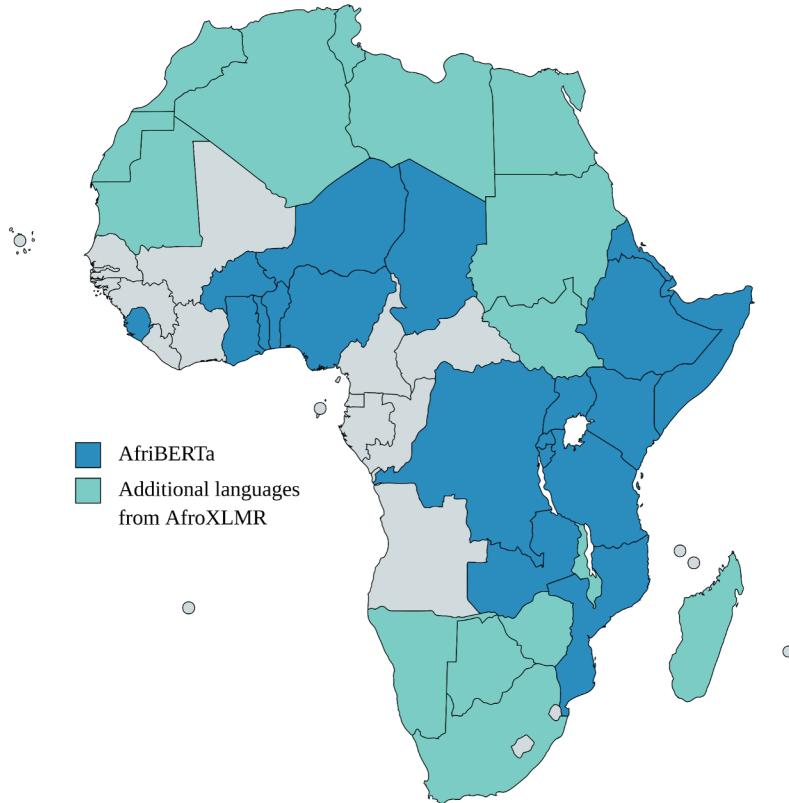
# Multilingual Adaptive Fine-tuning (MAFT)

MAFT To address limited language coverage



Tang et al. **Multilingual translation with extensible multilingual pretraining and finetuning**. arXiv 2020

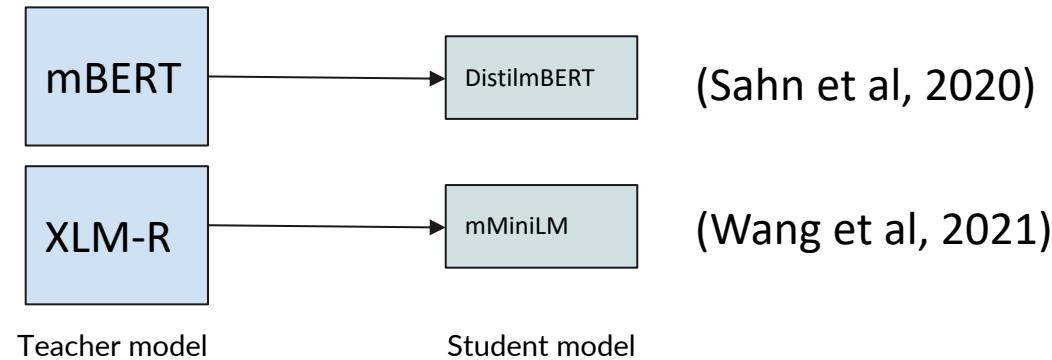
# MAFT for African languages



- **Choice of PLM**
  - **XLM-R-base**
  - Covers **Arabic, Latin and Ge'ez** scripts
- **Languages:**
  - English, French, Arabic, and
  - **17 most-resourced African languages (>50MB)**
- **Corpus**
  - mC4 – filtered web archive
  - News corpus

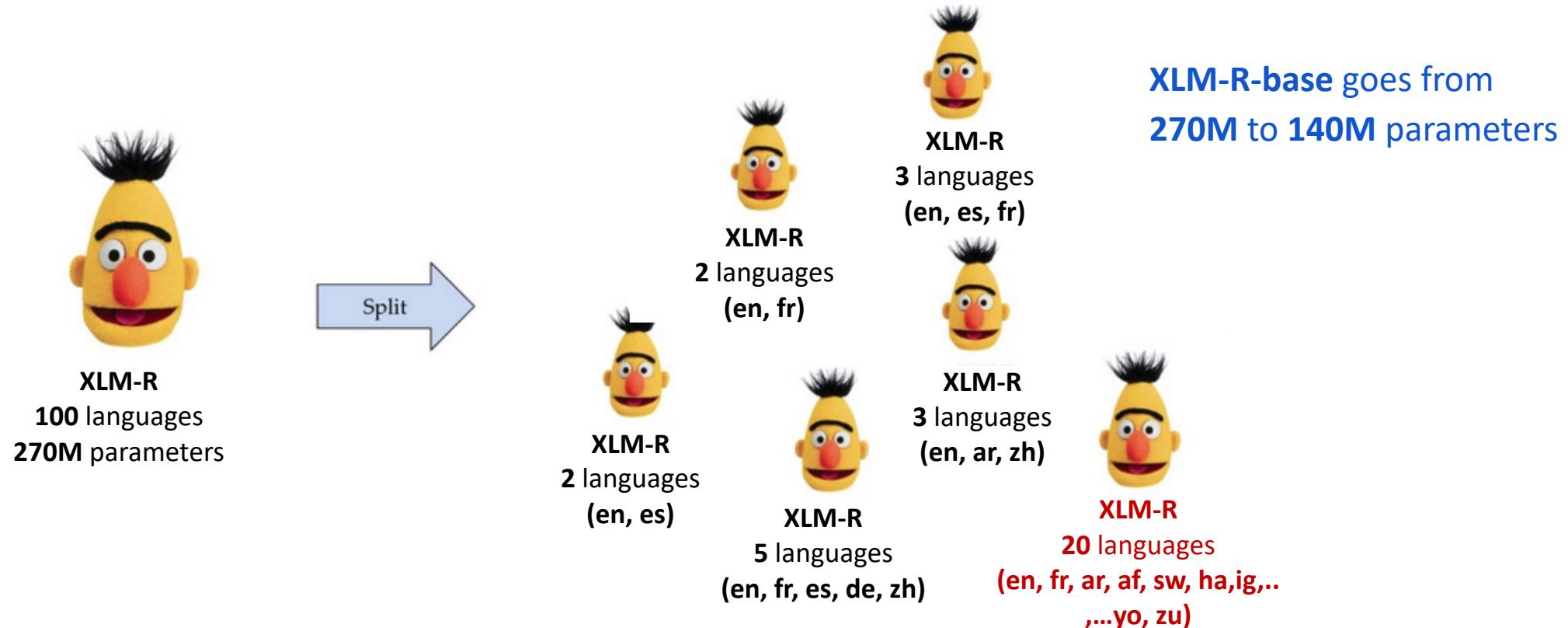
# Addressing “Hardware restrictions of PLMs”

- Multilingual pre-trained LMs (PLMs) are **quite big** (**>170 million parameters**)
  - Most labs in **low-resource communities** cannot run the models
  - **Fine-tuning also takes more time** for bigger models.
- **Knowledge distillation** comes to the rescue but *leads to significant drop* in performance

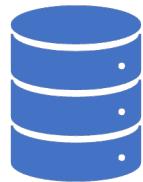


# Vocabulary reduction before MAFT

- An approach that leads to **competitive performance** compared to the **original multilingual PLM**
- **Vocab reduction + MAFT**



# Outline for AfroXLMR



Challenges of  
adapting PLM

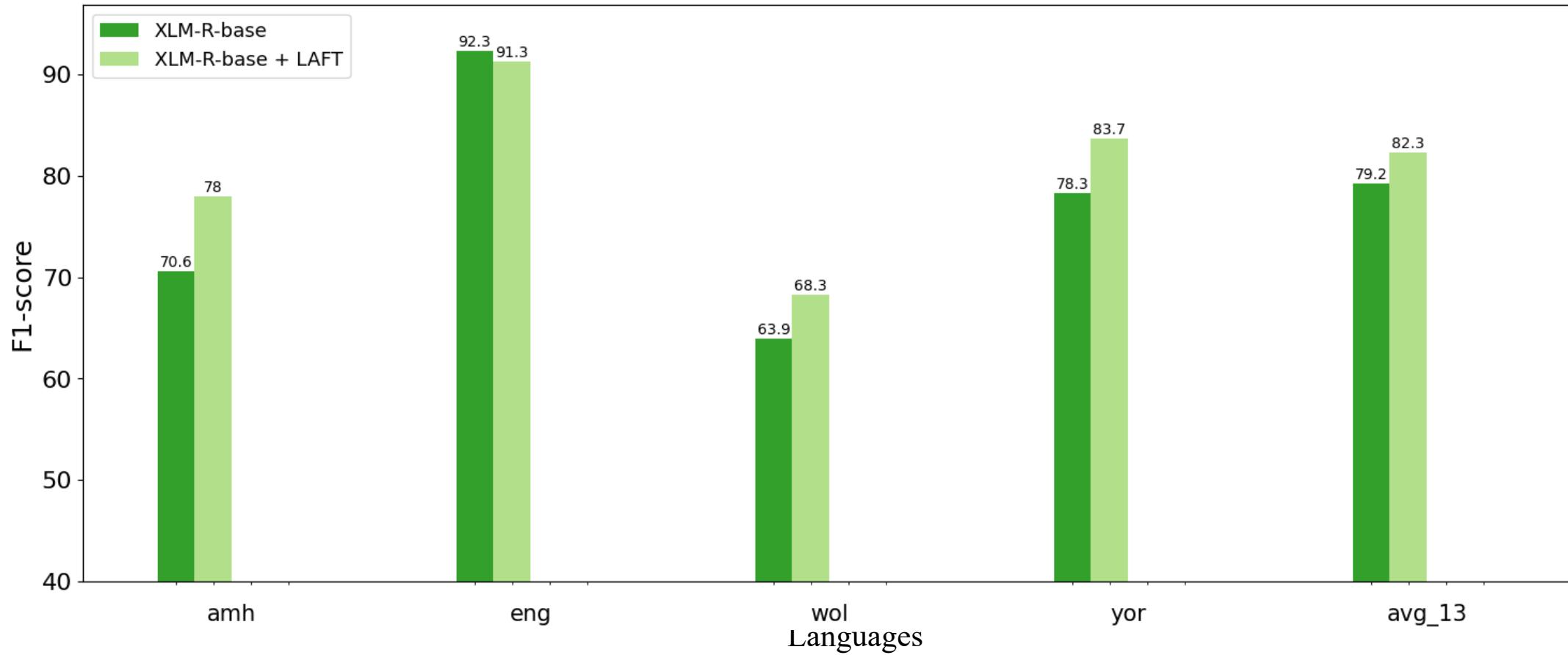


*Method*  
*Multilingual adaptive  
fine-tuning*



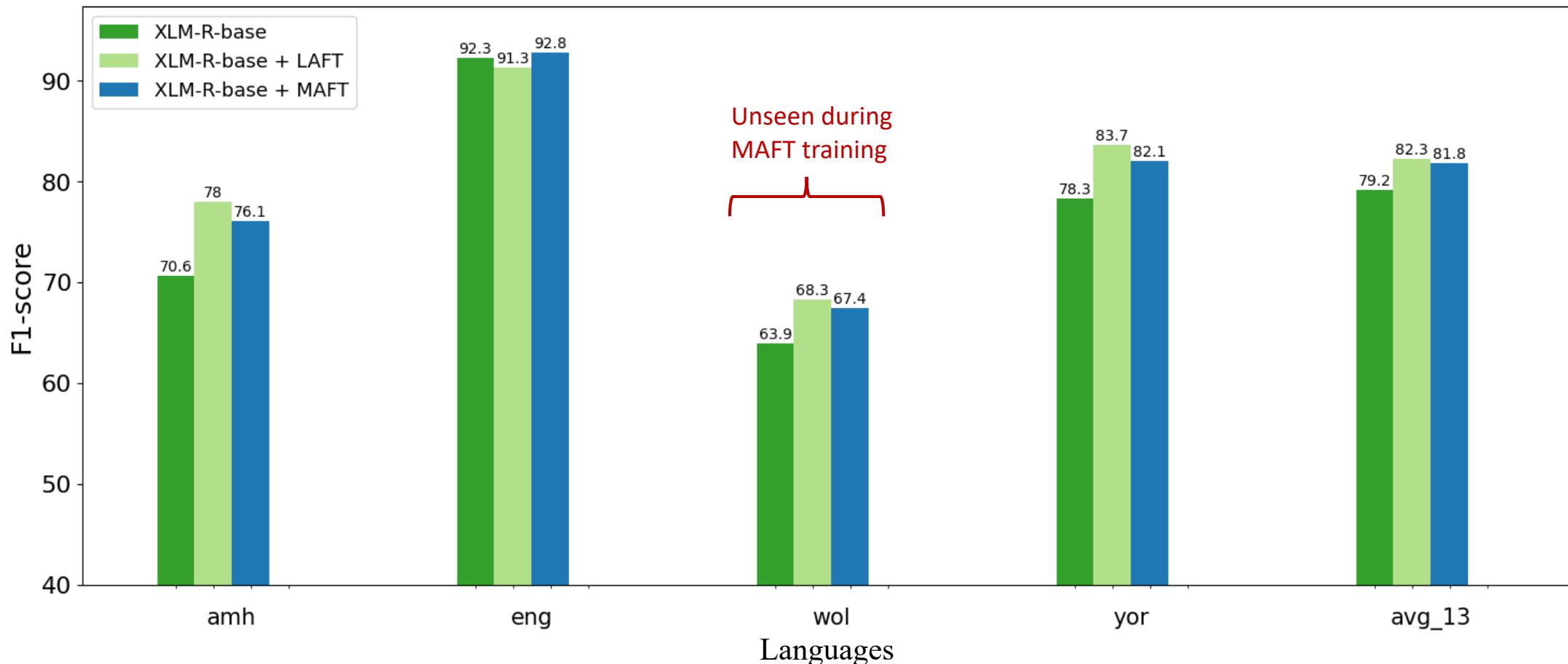
***Results***

# Evaluation of LAFT on MasakhaNER



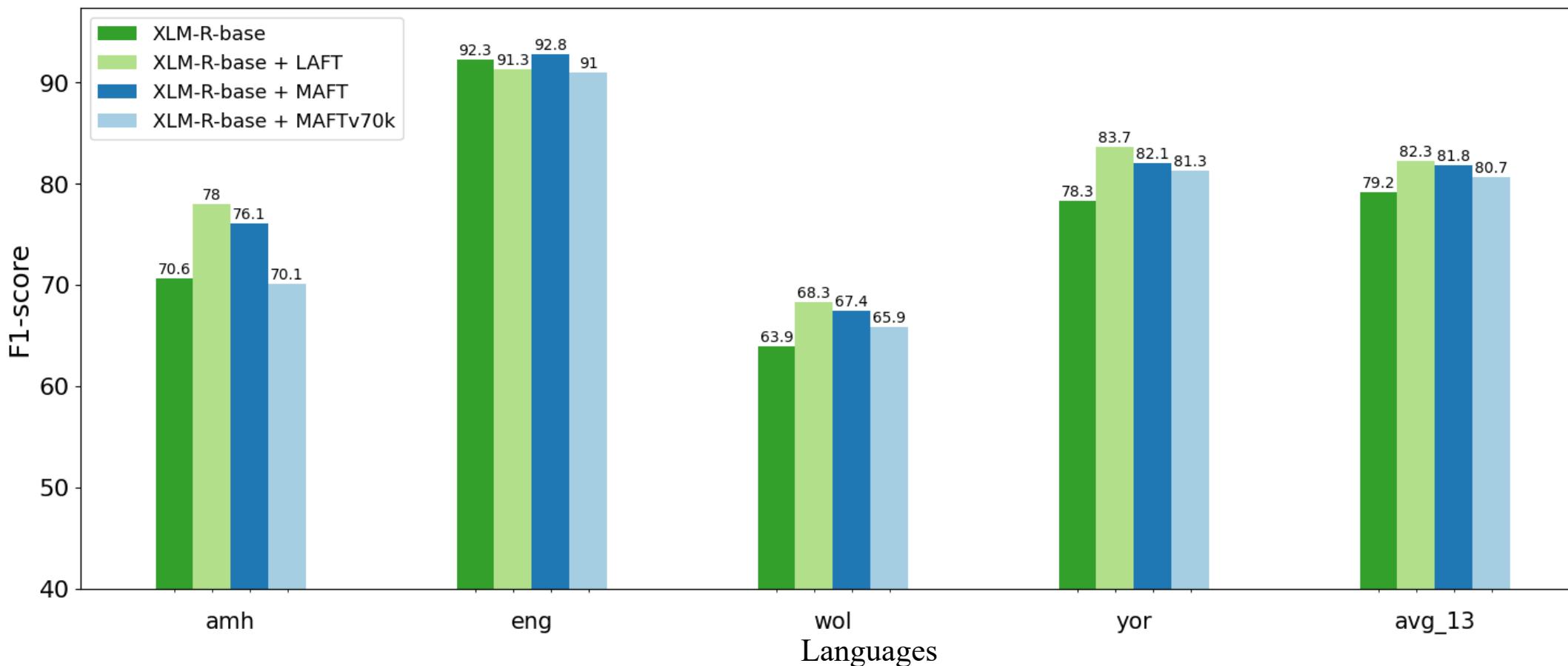
# Evaluation of MAFT on MasakhaNER

With MAFT, competitive performance and a single model

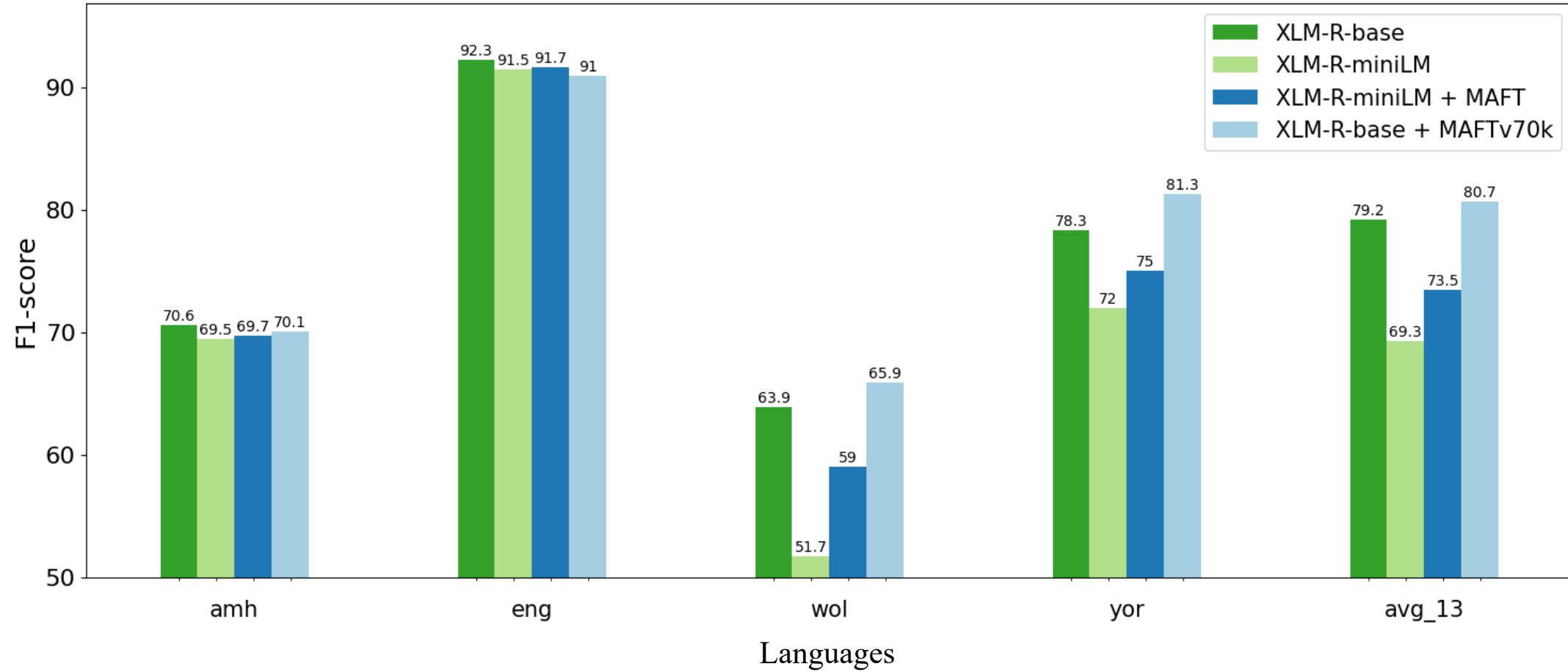


# Evaluation on MAFT - Vocab compression

Vocab reduction leads to slight drop in performance

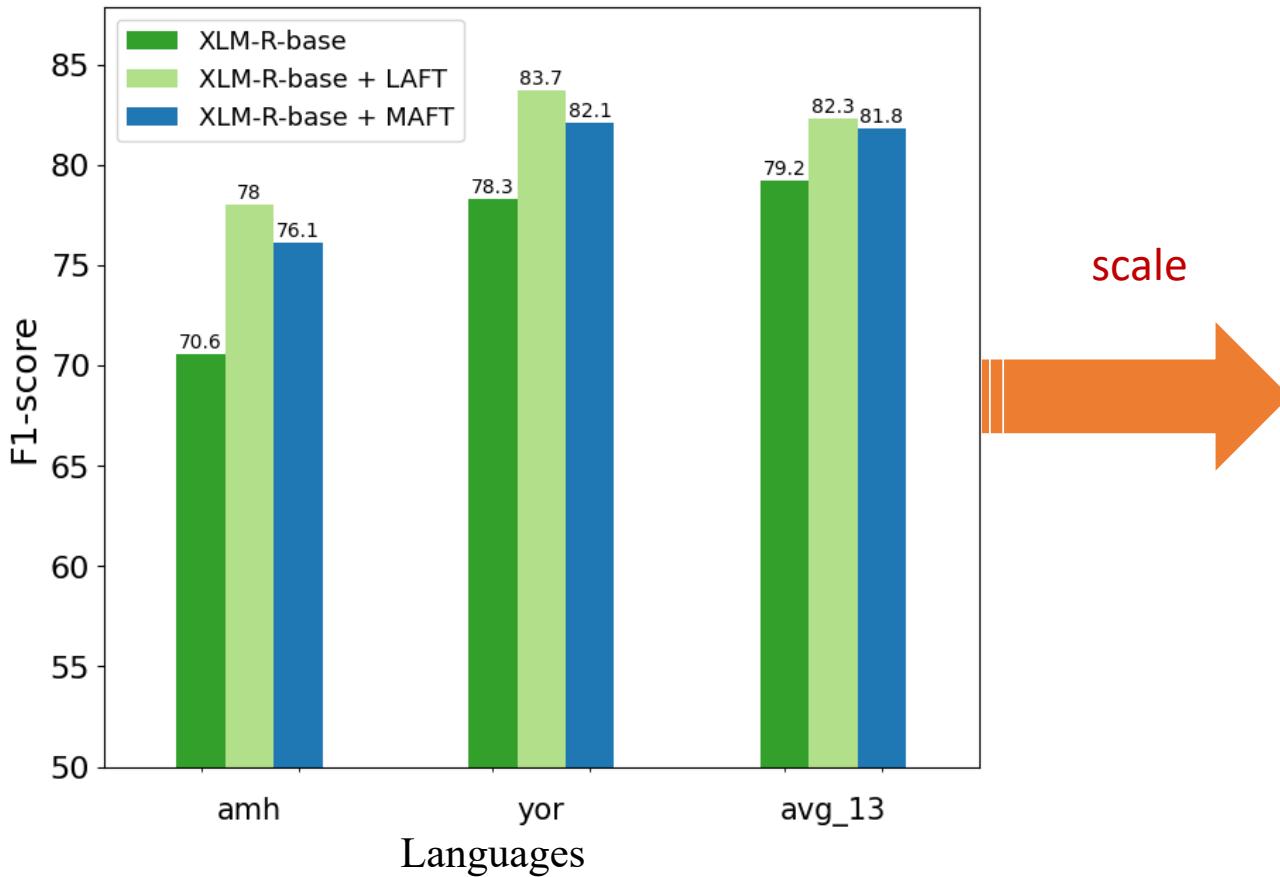


# Vocab compression is better than knowledge distillation

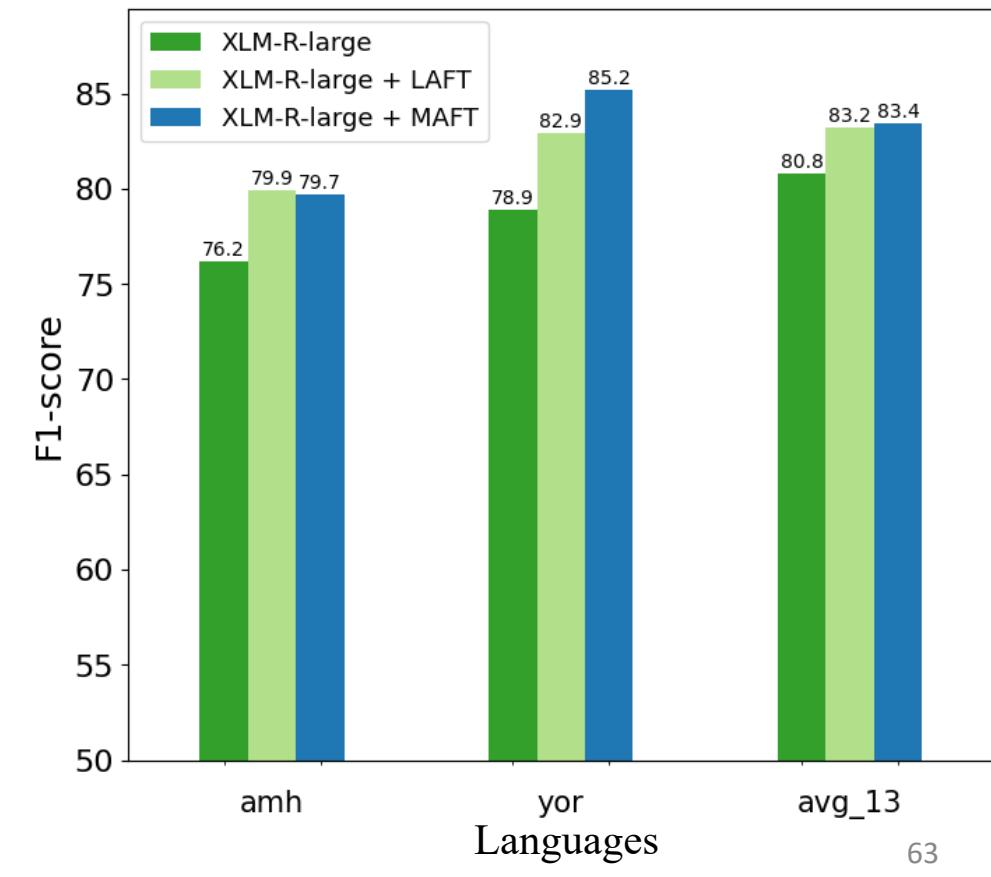


# MAFT on MasakhaNER with larger models

MAFT is either competitive or better than LAFT

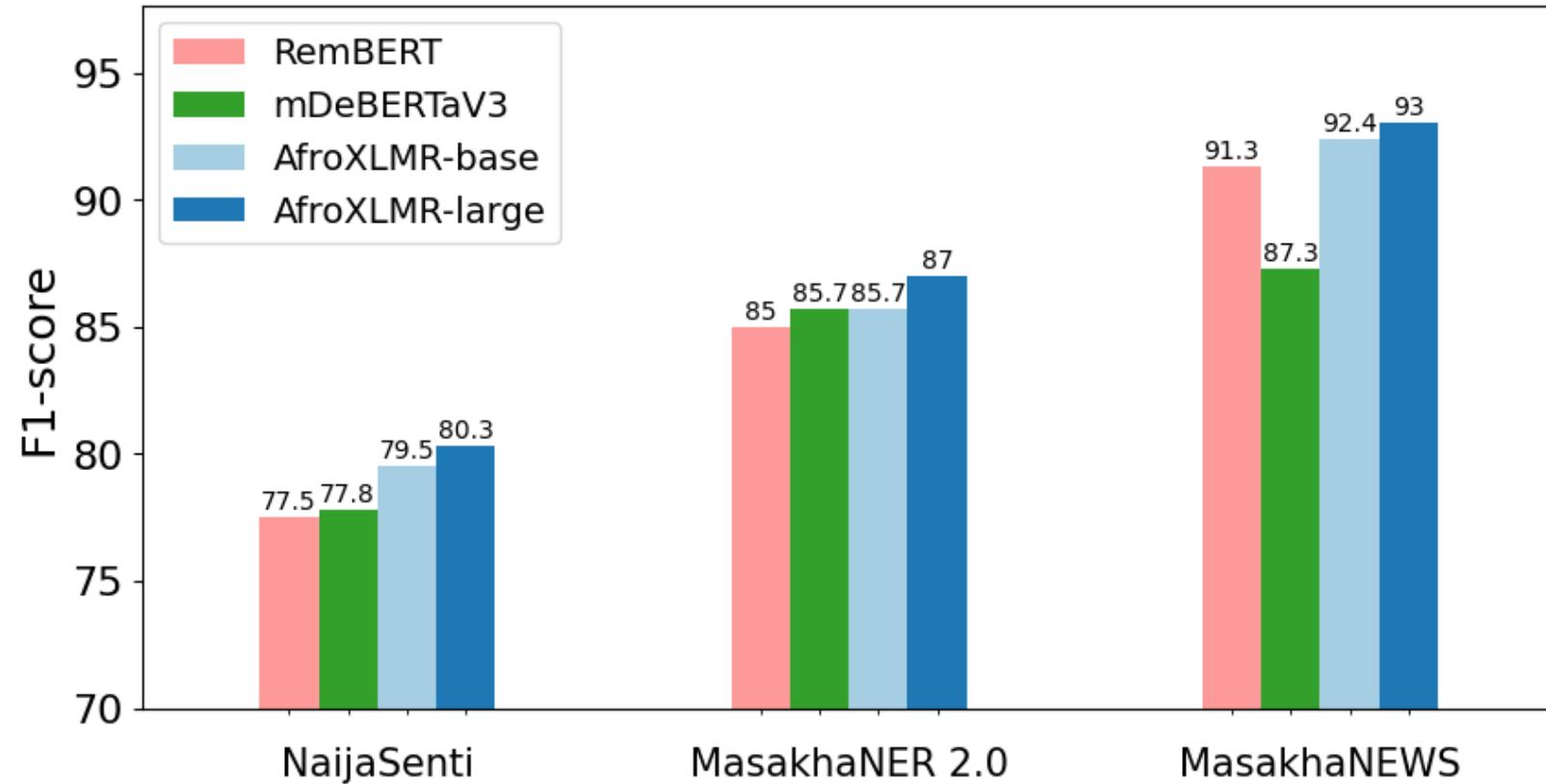


scale



# AfroXLMR achieves the state-of-the-art performance

- Comparing to **RemBERT** (559M parameters) and **mDeBERTaV3** (276M parameters)
  - **Sentiment classification (NaijaSenti), MasakhaNER and news topic classification (MasakhaNEWS)**



# AfroXLMR open-sourced on HuggingFace Model Hub

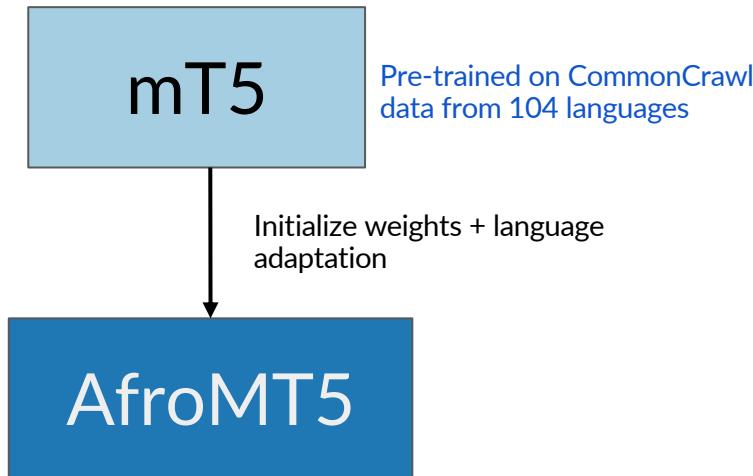
The screenshot shows the Hugging Face Model Hub interface for the **afro-xlmr-large** model. At the top, there is a search bar with the placeholder "Search models, datasets, users...". Below the search bar, the model's name **afro-xlmr-large** is displayed along with a profile picture of the owner, Davlan, and a "like" button showing 3 likes.

Below the name, several tags are listed: **Fill-Mask**, **PyTorch**, **Transformers**, **19 languages** (highlighted in green), **xlm-roberta**, **generated\_from\_trainer**, **AutoTrain Compatible**, and **License: mit**.

At the bottom of the card, there are buttons for **Train** (with a dropdown arrow), **Deploy** (with a dropdown arrow), and **Use in Transformers**. Below these buttons, there are links for **Model card**, **Files**, **Community**, and **Settings**.

A modal window titled **Edit model card** is open at the bottom of the card. Inside this modal, the model's name **afro-xlmr-large** is shown again, along with a section for **Downloads last month** which displays the value **14,215** next to a purple line graph showing the download trend over time.

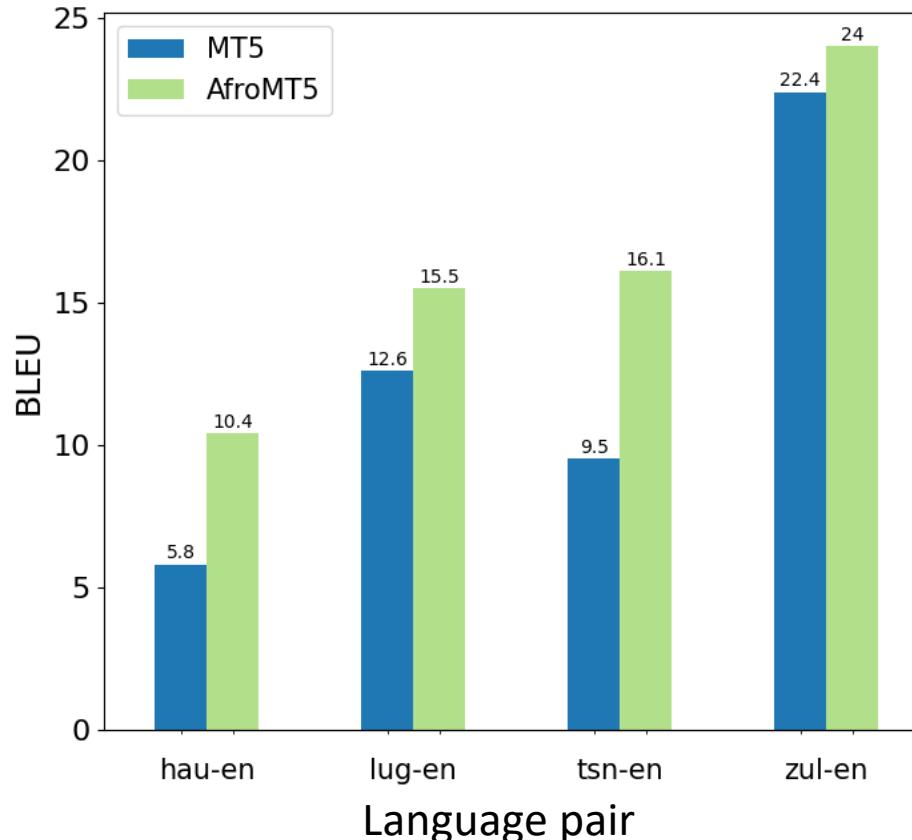
# MAFT for Sequence-to-Sequence Models



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

David Ifeoluwa Adelani<sup>1\*</sup>, Jesujoba Oluwadara Alabi<sup>2\*</sup>, Angela Fan<sup>3\*</sup>, Julia Kreutzer<sup>4\*</sup>, Xiaoyu Shen<sup>5</sup>, Machel Reid<sup>6\*</sup>, Dana Ruiter<sup>1</sup>, Dietrich Klakow<sup>1</sup>, Peter Nabende<sup>7\*</sup>, Ernie Chang<sup>1\*</sup>, Tajuddeen R. Gwadabe<sup>8\*</sup>, Freshia Sackey<sup>9\*</sup>, Bonaventure F. P. Dossou<sup>10\*</sup>, Chris Chinene Emezue<sup>11\*</sup>, Colin Leong<sup>12\*</sup>, Michael Beukman<sup>13\*</sup>, Shamsuddeen H. Muhammad<sup>14\*</sup>, Guyo D. Jarso<sup>\*</sup>, Oreen Yousuf<sup>15\*</sup>, Andre N. Rubungo<sup>16\*</sup>, Gilles Hacheme<sup>17\*</sup>, Eric P. Wairagala<sup>7\*</sup>, Muhammad U. Nasir<sup>18\*</sup>, Benjamin A. Ajibade<sup>\*</sup>, Tunde Oluwaseyi Ajayi<sup>\*</sup>, Yvonne Wambui Gitau<sup>\*</sup>, Jade Abbott<sup>\*</sup>, Mohamed Ahmed<sup>19\*</sup>, Millicent Ochieng<sup>19\*</sup>, Anuoluwapo Aremu<sup>\*</sup>, Perez Ogayo<sup>20\*</sup>, Jonathan Mukibi<sup>7\*</sup>, Fatoumata Ouoba Kabore<sup>\*</sup>, Godson Koffi Kalipe<sup>\*</sup>, Derguene Mbaye<sup>21\*</sup>, Allahsera Auguste Tapo<sup>22\*</sup>, Victoire M. Koagne<sup>\*</sup>, Edwin Munkoh-Buabeng<sup>\*</sup>, Valencia Wagner<sup>23\*</sup>, Idris Abdumumin<sup>24\*</sup>, Ayodele Awokoya<sup>25\*</sup>, Happy Buzaaba<sup>\*</sup>, Blessing Sibanda<sup>26\*</sup>, Andiswa Bukula<sup>27\*</sup>, Sam Manthalu<sup>28</sup>

Leveraging **pre-trained models** using a **few thousand parallel sentences** to build a machine translation model for a **new domain**



# How to adapt an English LLM to new language (s)?

# Language adaptation for Large Language Models

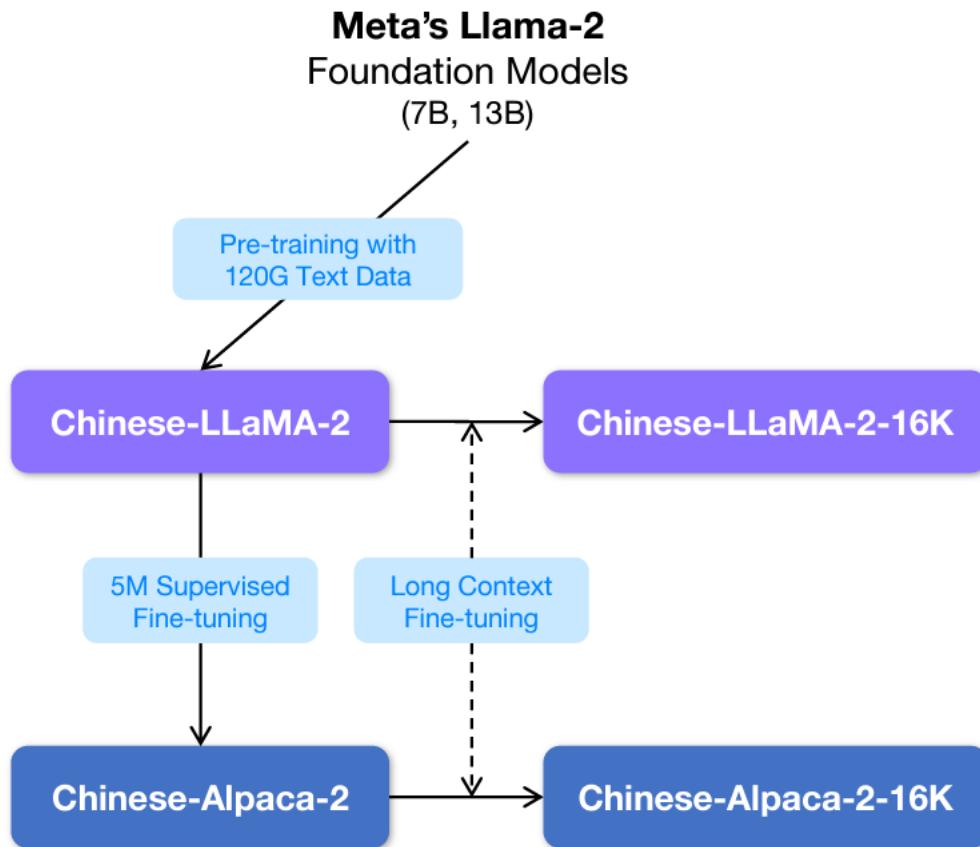
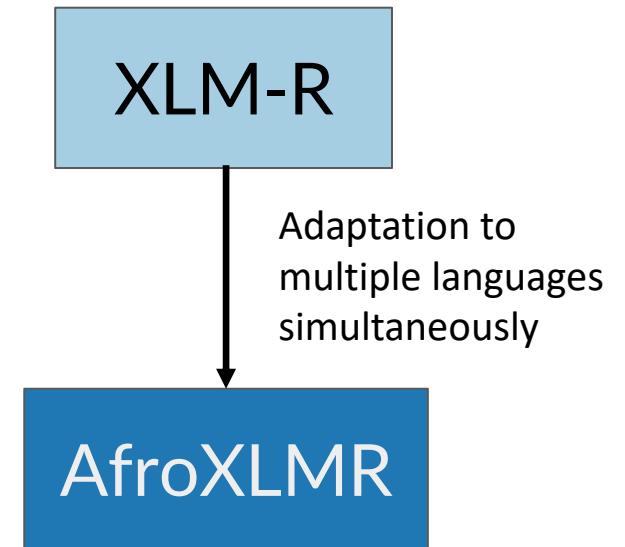


Table 11: Results on CMMLU test sets.

Model	Test Set	
	Zero-shot	Few-shot
Chinese-LLaMA-2-7B	27.9	34.1
Chinese-LLaMA-2-13B	38.9	42.5
Chinese-Alpaca-2-7B	40.0	41.8
Chinese-Alpaca-2-13B	43.2	45.5

# Summary of Part 2

- We tackle **two major challenges of adapting pre-trained language models (PLMs)** to under-resourced languages:
  - Limited language coverage in PLMs
  - Hardware restrictions of fine-tuning PLMs
- We propose **multilingual adaptive fine-tuning (MAFT)** to address the first challenge.
- We propose **vocabulary reduction of the PLM followed by MAFT** to address the second challenge.



# PART THREE

**Challenge 3: Choosing the best transfer language**



**Approach to address it: LangRank with co-training and PEFT**

## Cross-lingual transfer learning for NER and POS

**MasakhaNER 2.0** (Adelani et al, EMNLP 2022)

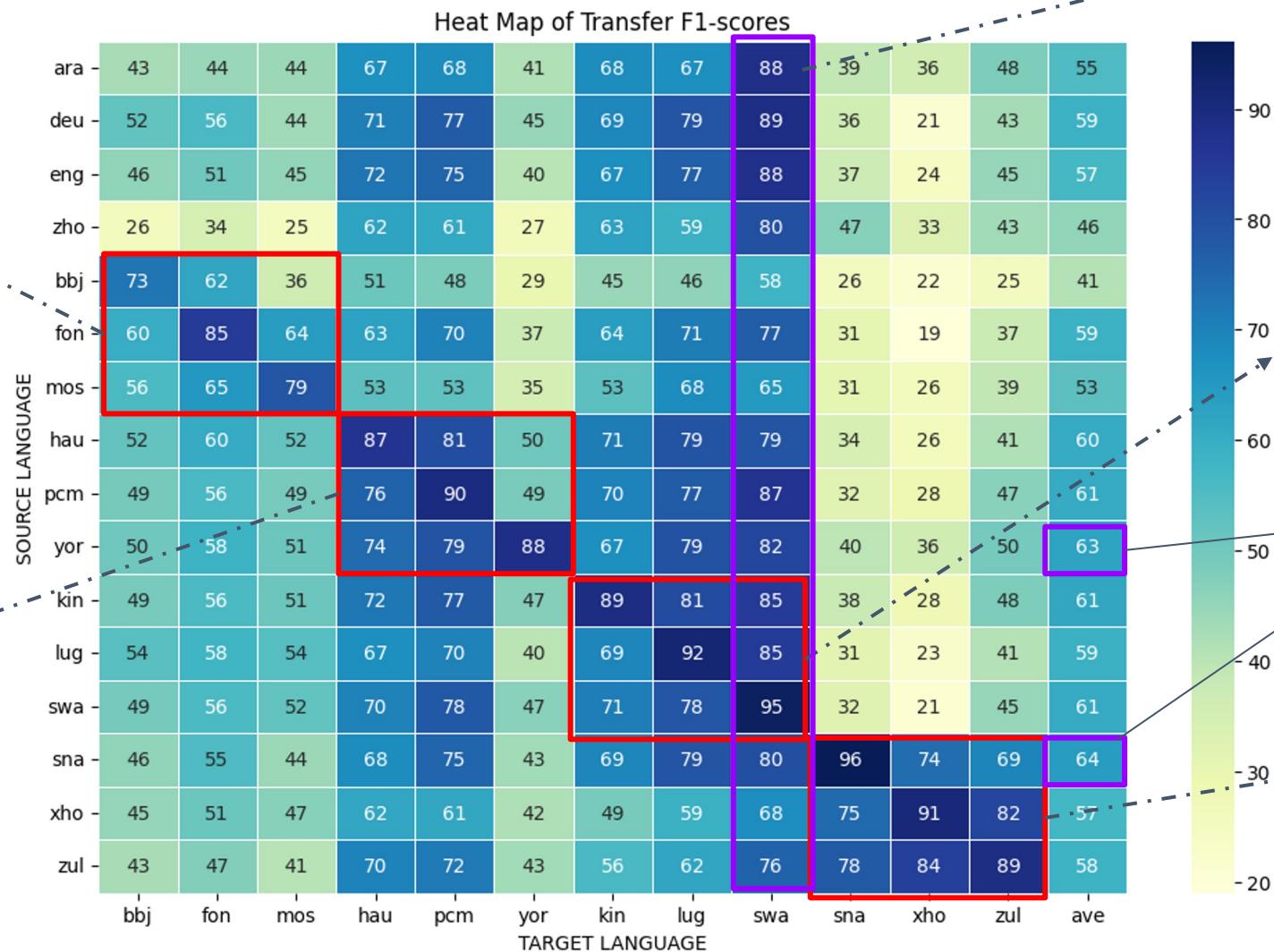
**MasakhaPOS** (Dione & Adelani et al., ACL 2023)

# Brute-force approach on choosing transfer languages.

- For each **source** language
  - Train on source language
  - Evaluate on all available target languages.
- Suppose we have **N**-languages
  - $N \times N$  matrix of transfer scores.
- We collected **42 NER datasets online**
  - 21 from non-African languages
  - 21 from African (MasakhaNER dataset)
  - We evaluate  $42 \times 42$  transfer scores.

# Heat map of transfer scores

Geographical proximity and entity overlap in Francophone Africa



Geographical proximity among Nigerian languages and entity overlap in Nigeria

Best transfer language may not always be African. Swahili benefits from Arabic due to trade influence despite difference in script

Swahili is easy to transfer to in general

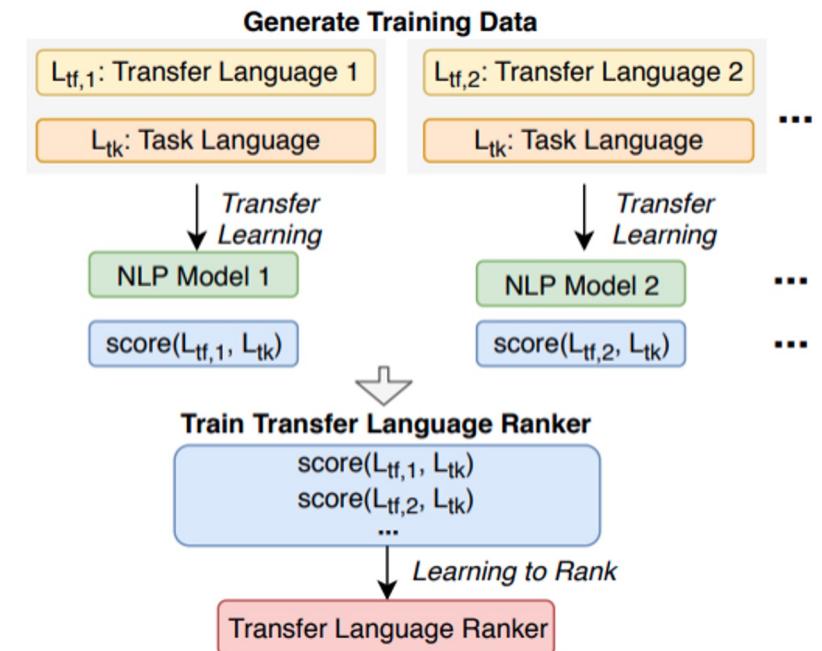
Syntactic similarity and entity overlap

chiShona and Yoruba are best transfer languages on average

Syntactic similarity (chiShona, isiXhosa, and isiZulu), genealogical closeness, and entity overlap

# Does this approach scale to new languages?

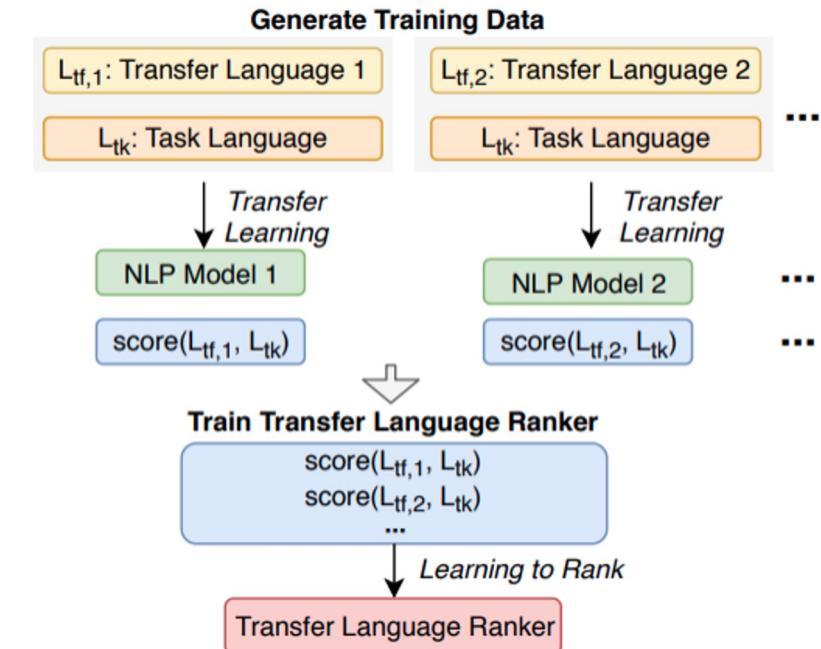
- No - need to repeat this for a new language.
- What if the new language do not have evaluation data?
- Lin et al. (2019) proposed **LangRank**
  - Train a set of NLP models with all available transfer languages and collect evaluation scores.
  - Train a ranking model to predict the top transfer languages
    - Ranking model is trained using evaluation score, and several linguistic and data features.
  - Predict transfer languages for a new language.



Workflow of learning to select the transfer language

# LangRank features

- Linguistic distance measures from lang2vec
  - Geographic distance (**d\_geo**)
  - Genetic distance (**d\_gen**)
  - Inventory distance (**d\_inv**)
  - Syntactic distance (**d\_syn**)
  - Phonological distance (**d\_pho**)
  - Featural distance (**d\_fea**)
- Data-dependent features:
  - Transfer language dataset size (**s\_tf**)
  - Target language dataset size (**s\_tg**)
  - Transfer over target size ratio (**sr**)
  - Entity Overlap (**eo**)
- Source-target transfer evaluation score



We make use of a **boosting based ranker** trained using **Leave-out-one approach**, excluding the target language.

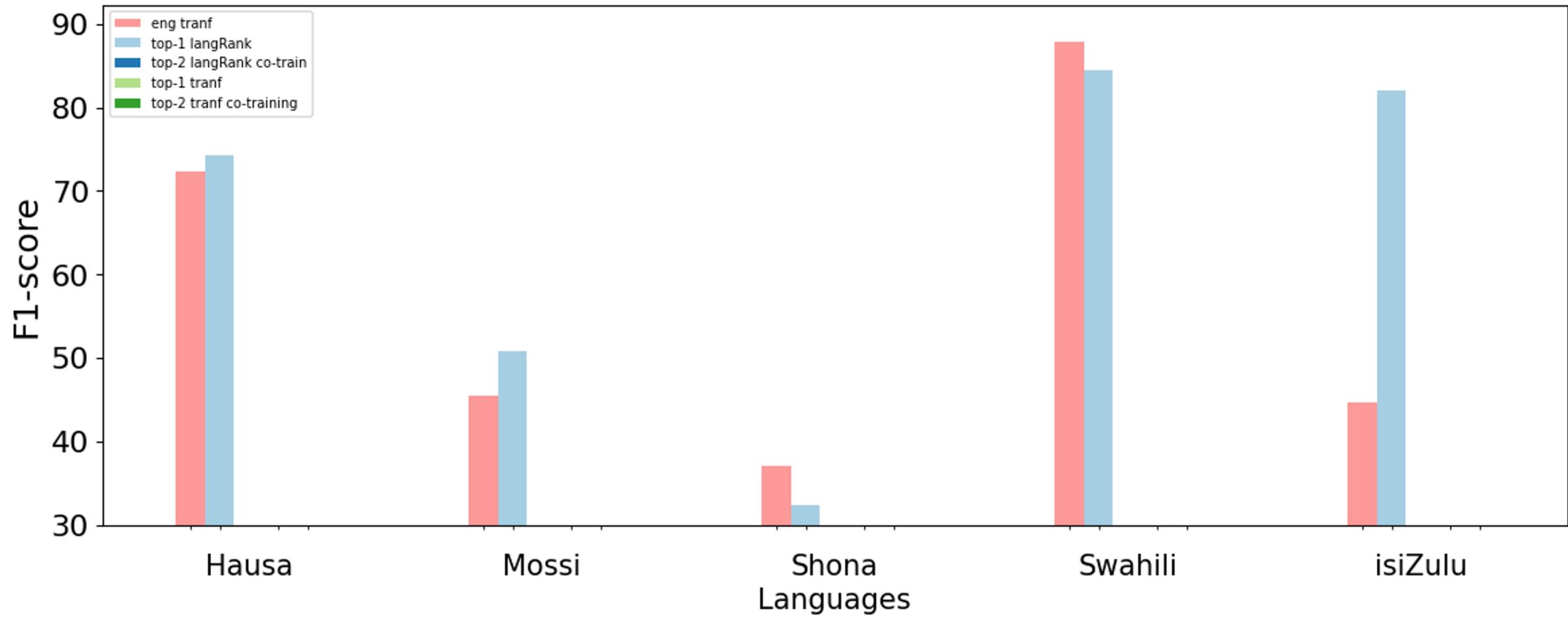
# LangRank selected languages and features

- LangRank predicted the right language as **one of the top-2 best transfer language in 13 target languages**
- For the remaining languages, the **alternative provided by LangRank is not worse than -5 points** in F1.

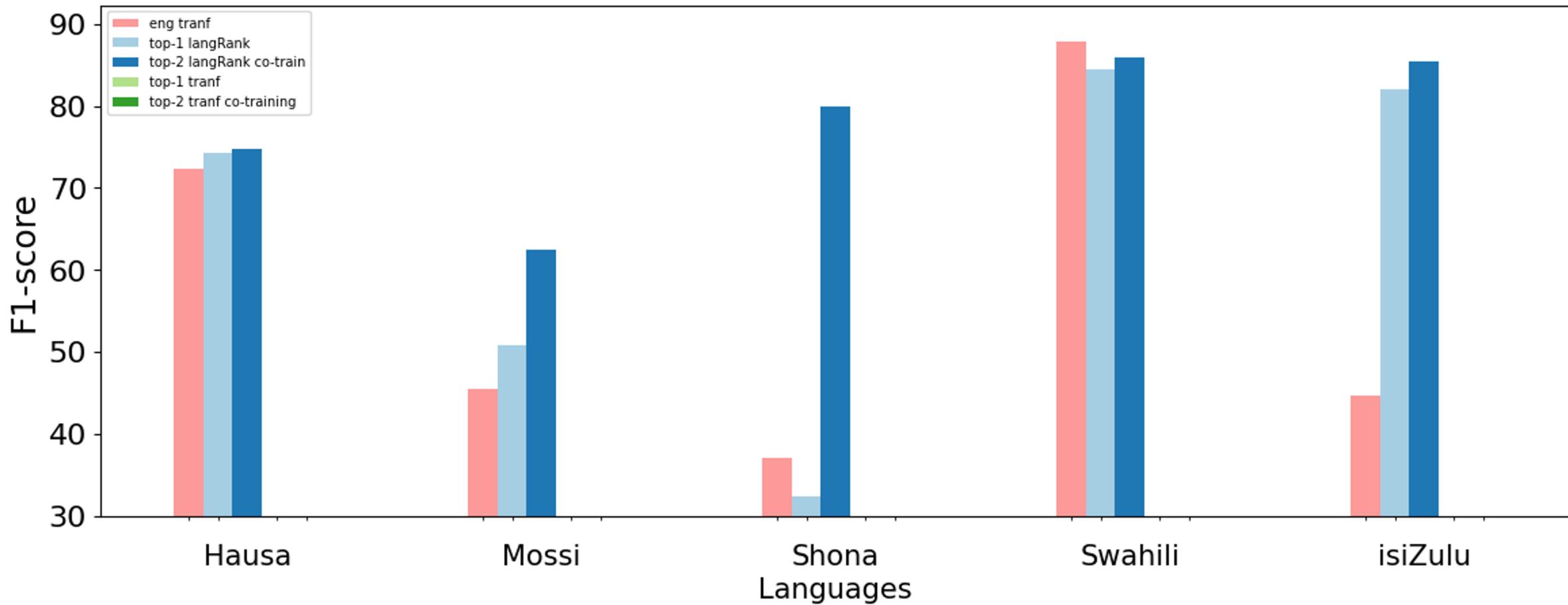
Target Lang.	Top-2 Transf. Lang	Top-2 LangRank Model	Top-3 features selected by LangRank model Lang 1; Lang 2
bbj	fon, ewe	twi, ewe	( $s_{tf}$ , $d_{syn}$ , $d_{geo}$ ); ( $s_{tf}$ , $d_{geo}$ , $sr$ )
fon	mos, bbj	yor, ewe	( $d_{geo}$ , $d_{syn}$ , $sr$ ); ( $s_{tf}$ , $d_{geo}$ , $d_{gen}$ )
hau	pcm, yor	yor, swa	( $d_{geo}$ , $sr$ , $eo$ ); ( $eo$ , $sr$ , $s_{tf}$ )
mos	fon, ewe	yor, fon	( $d_{geo}$ , $d_{inv}$ , $sr$ ); ( $d_{geo}$ , $s_{tf}$ , $sr$ )
pcm	hau, yor	eng, yor	( $eo$ , $d_{gen}$ , $d_{syn}$ ); ( $eo$ , $d_{geo}$ , $sr$ )
sna	zul, xho	swa, zul	( $eo$ , $sr$ , $s_{tf}$ ); ( $d_{geo}$ , $sr$ , $eo$ )
swa	deu, ara	ita, nld	( $sr$ , $d_{inv}$ , $eo$ ); ( $eo$ , $s_{tf}$ , $sr$ )
xho	zul, sna	zul, pcm	( $eo$ , $d_{geo}$ , $d_{gen}$ ); ( $eo$ , $s_{tf}$ , $d_{inv}$ )
yor	hau, pcm	fon, pcm	( $d_{geo}$ , $d_{inv}$ , $d_{syn}$ ); ( $eo$ , $d_{geo}$ , $d_{inv}$ )
zul	xho, sna	xho, sna	( $eo$ , $d_{gen}$ , $d_{geo}$ ); ( $d_{syn}$ , $sr$ , $d_{geo}$ )

Most important features are  
**geographical distance (d\_geo)** and  
**entity overlap (eo)**

# Brute-force vs. LangRank results

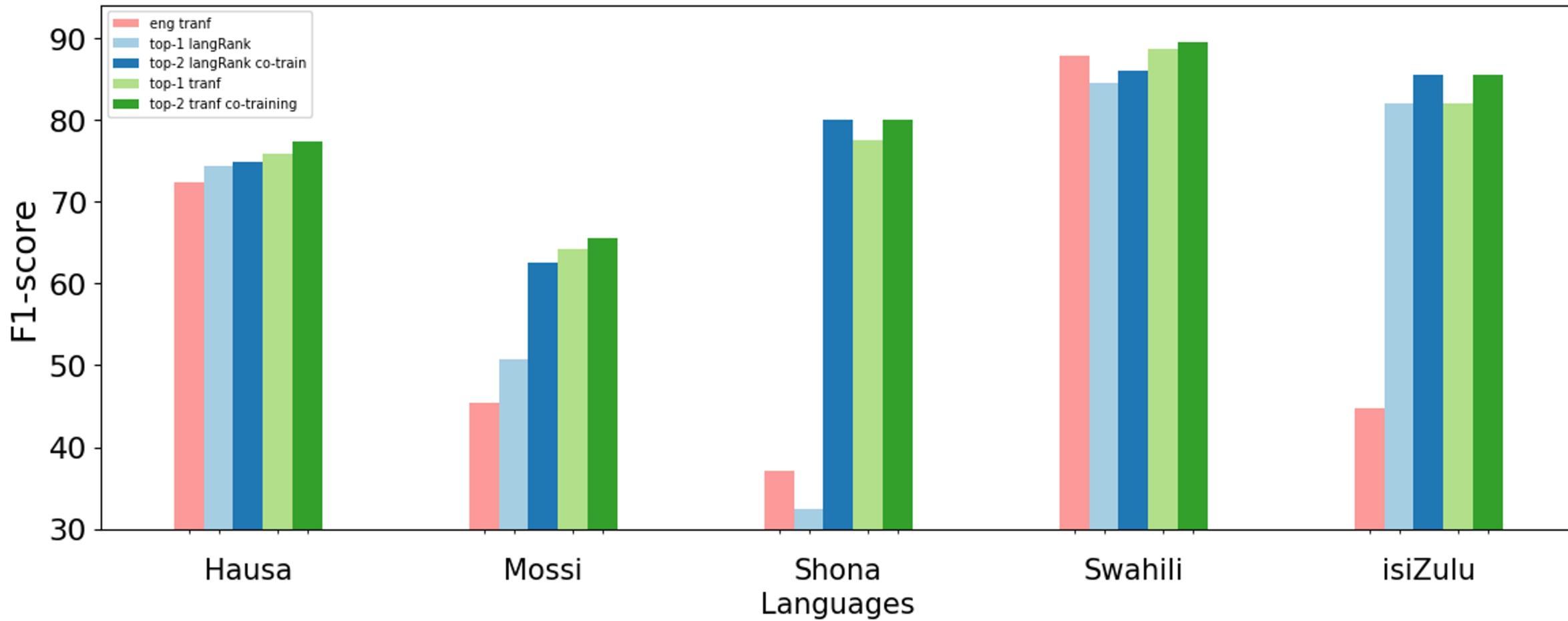


# Co-training top-2 best transfer languages help



# Co-training top-2 best transfer languages help

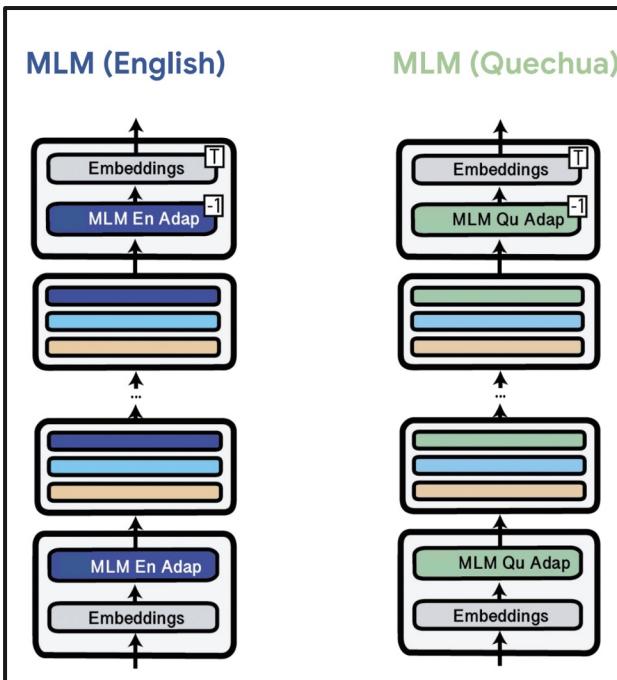
Performance is similar to the true best transfer language.



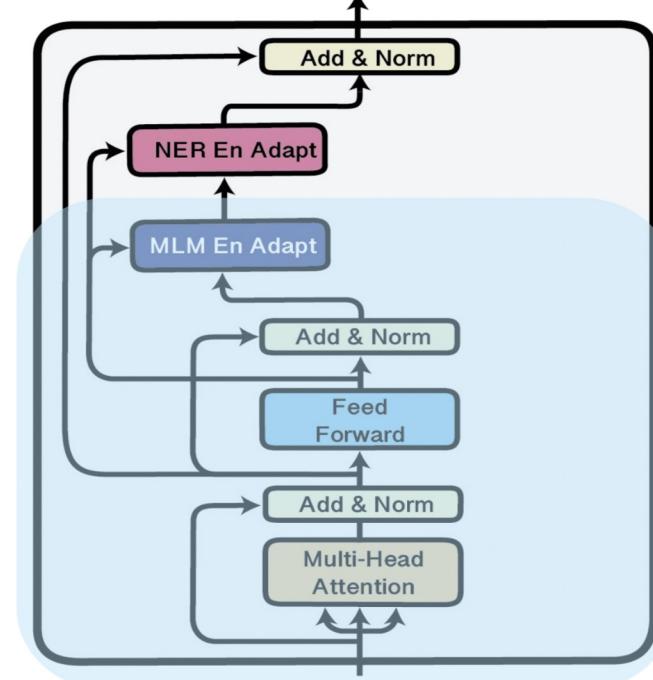
# Leveraging target language monolingual data with PEFT and best transfer language

Parameter-efficient fine-tuning with adapters: MAD-X approach

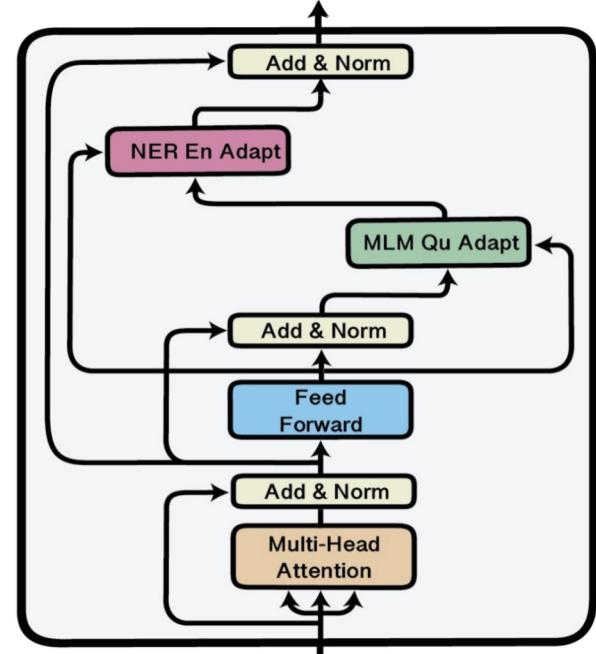
**STEP 1: Train language adapters on every language of interest**



**STEP 2: Train task adapter together with the language adapter, only modify task adapter, others parameters are frozen**

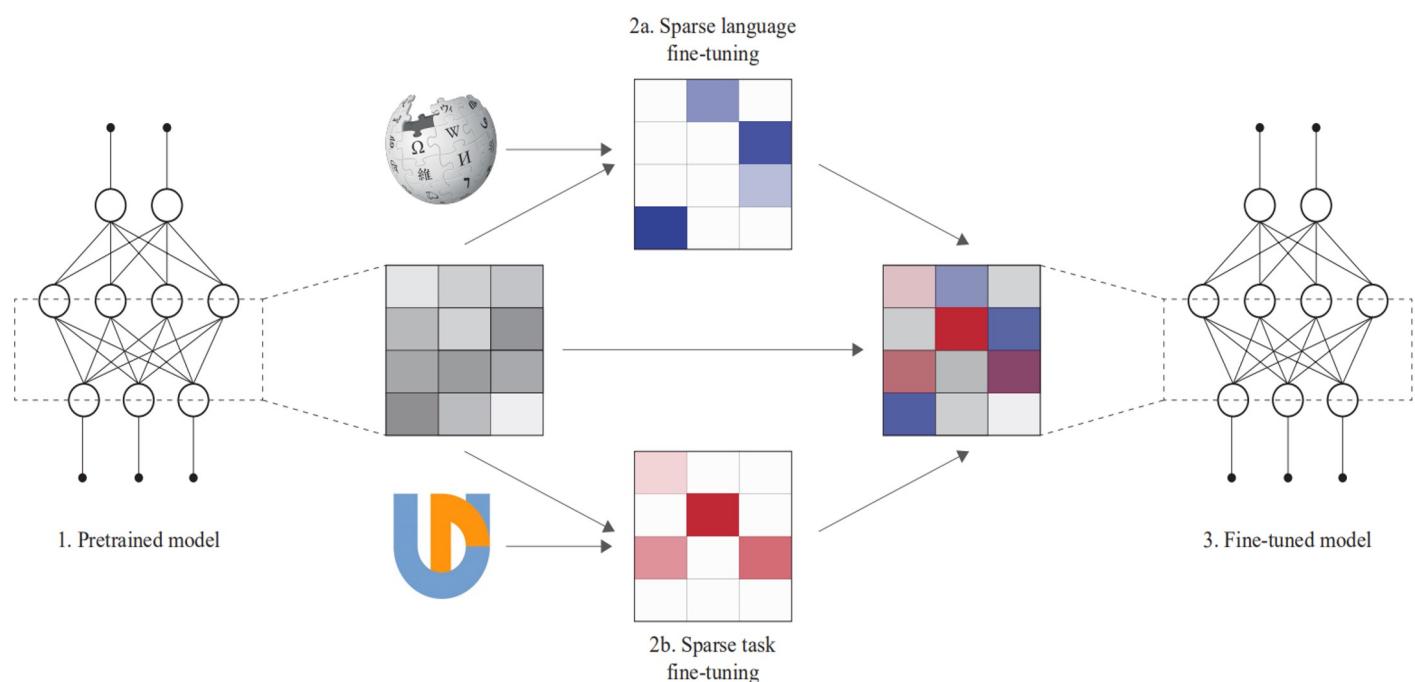


**STEP 3: Zero-Shot Transfer to Target Language**  
by replacing the source language adapter but keeping the task adapter.



# Lottery Ticket Sparse Fine-tunings (LT-SFT)

**Lottery Ticket Hypothesis (LTH)** (Frankle & Carbin, 2019) that states that **each neural model contains a sub-network (a "winning ticket")** that, *if trained again in isolation, can reach or even surpass the performance of the original model.*



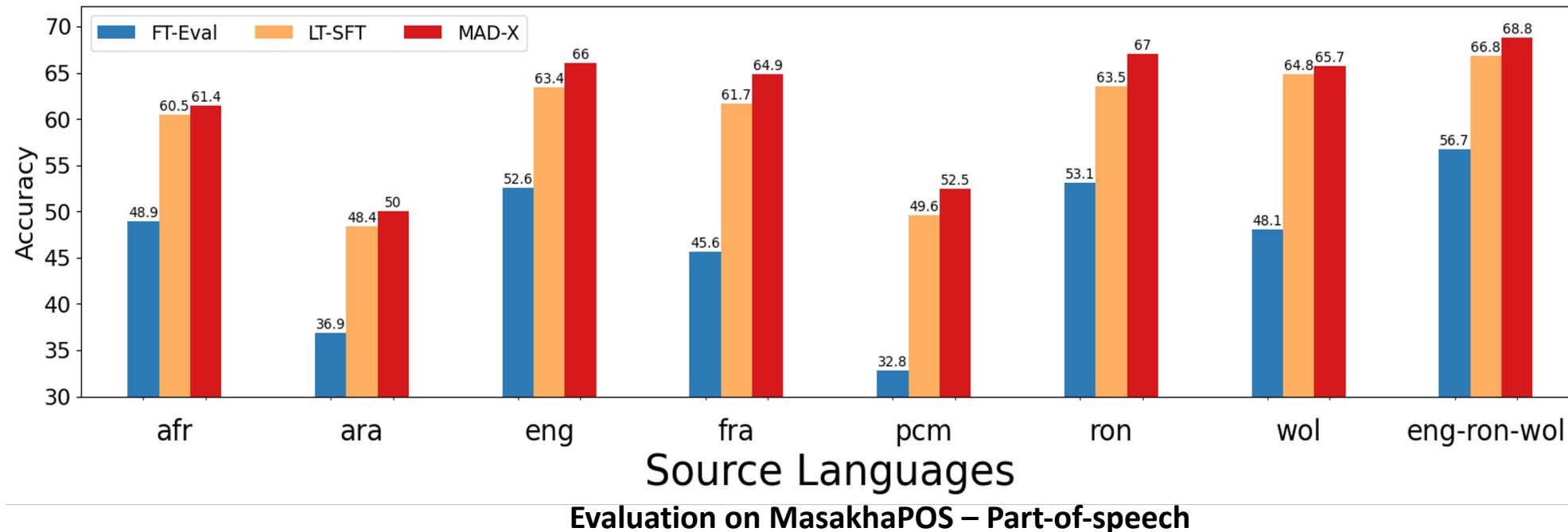
# Results on POS

Geographically close source languages: Afrikaans, Arabic, English, French, Naija, Romanian and Wolof

FT-Eval: Fine-tune PLM on a source language, perform zero-shot evaluation on the target language.

Parameter-efficient fine-tuning methods:

- LT-SFT (Ansell et al., 2021)
- MAD-X (Pfeiffer et al, 2020)



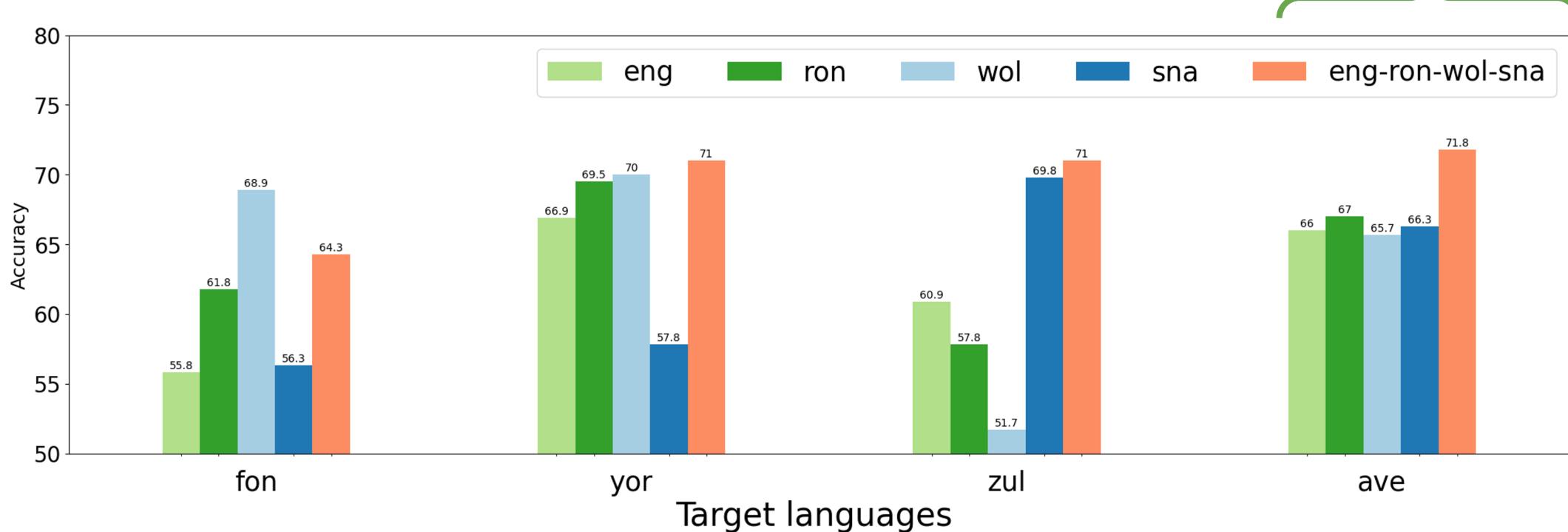
# More fine-grained results for POS

How important is the **source language** using MAD-X adaptation?

- English, Romanian, Wolof, and Shona

Evaluation on: Fon, Yoruba, and isiZulu

Multi-source/co-training  
adaptation is very effective



# Summary of Part 3

- Choosing the best transfer languages is often better than using English language dataset or any available data.
- When there are several options, better to choose carefully the best source languages
- We demonstrated this on NER and POS tasks.

# Thank you

**Challenge 1:** Lack of labelled data for downstream tasks

**Approach to address it:** Developing labelled datasets using participatory approach

**Challenge 2:** Limited coverage of PLM for new languages and scripts

**Approach to address it:** Developing multilingual language models through Multilingual adaptive fine-tuning

**Challenge 3:** Choosing the best transfer language

**Approach to address it:** LangRank with co-training and PEFT