Multilinguality:

How can we train Llama 5 200B on 100 Languages?

CSE 5525: Foundations of Speech and Natural Language Processing

https://shocheen.github.io/cse-5525-fall-2025/

THE OHIO STATE UNIVERSITY

Slide Credits: Greg Durrett, Ana Marasović, Antonios Anastasopoulos

Agenda

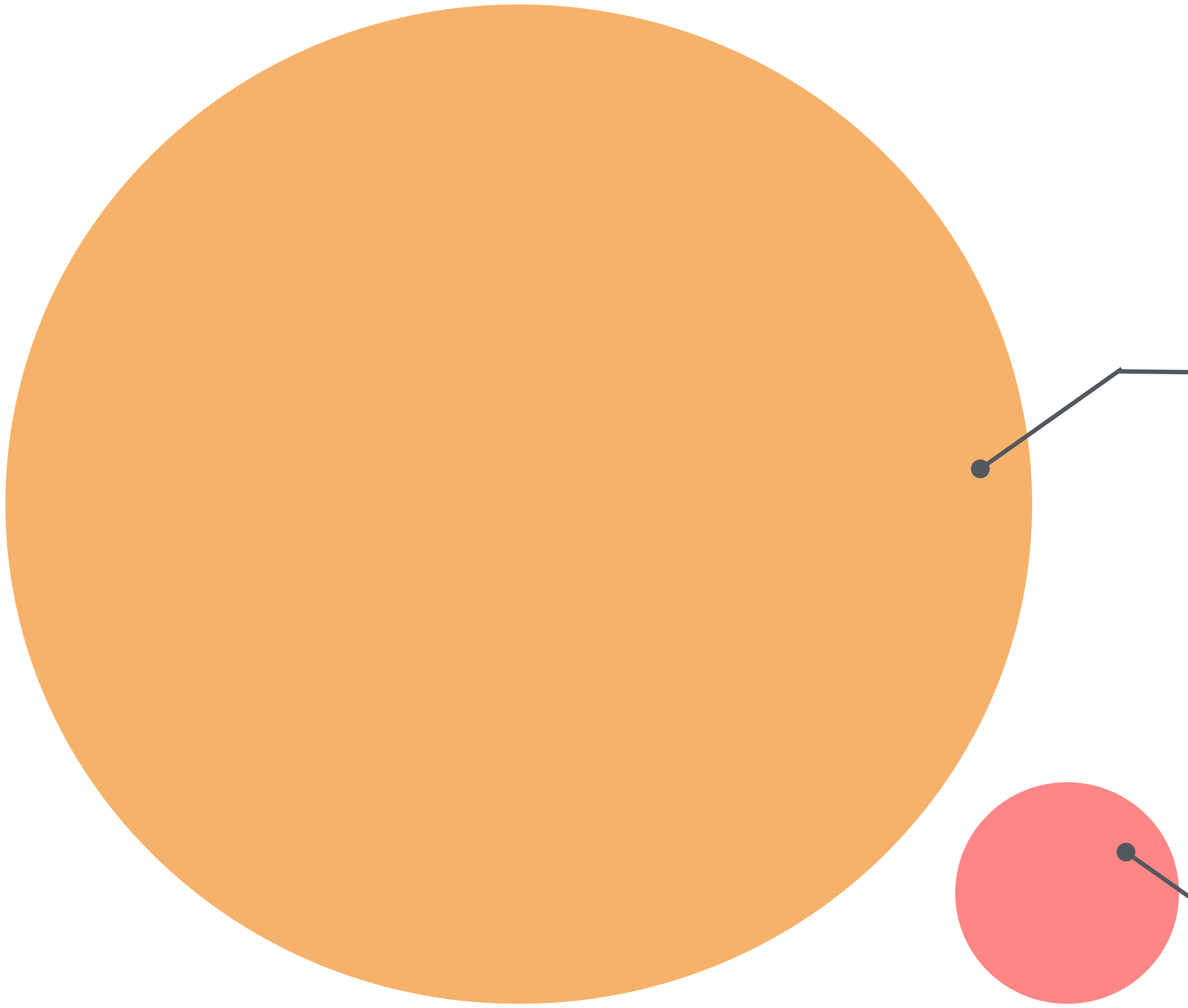
I. Languages of the World and Linguistic Diversity
II. Multilingual LLMs
   A. Pre-training
   B. Instruction Fine-tuning
   C. Alignment
III. Challenges
IV. Other Directions

Languages of the World

Image from [Ethnologue's website](#) ("Eberhard, et al.,. 2024. Ethnologue: Languages of the World. Twenty-seventh edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.")

7164 languages
in use !!! (as of 2022)

Languages by Vitality

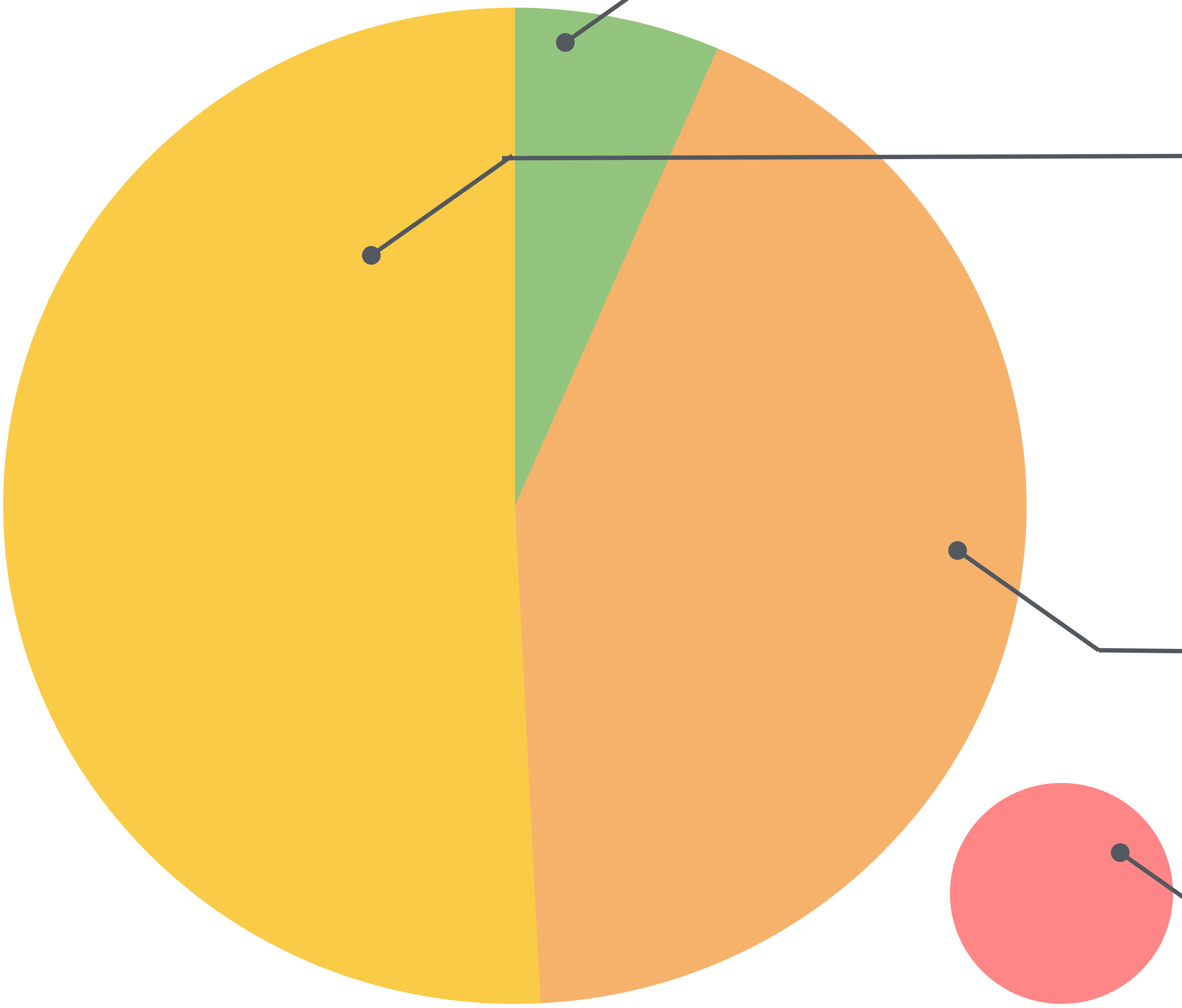


~7100 living languages

451 **extinct** languages:
Languages no longer used; no one identifies with the language

Image derived from the language vitality data from Ethnologue's website ("Eberhard, et al., 2024. Ethnologue: Languages of the World. Twenty-seventh edition. Dallas, Texas: SIL International.). Image approx. at scale.

Languages by Vitality



492 **institutional** languages:
Sustained by institutions / governments

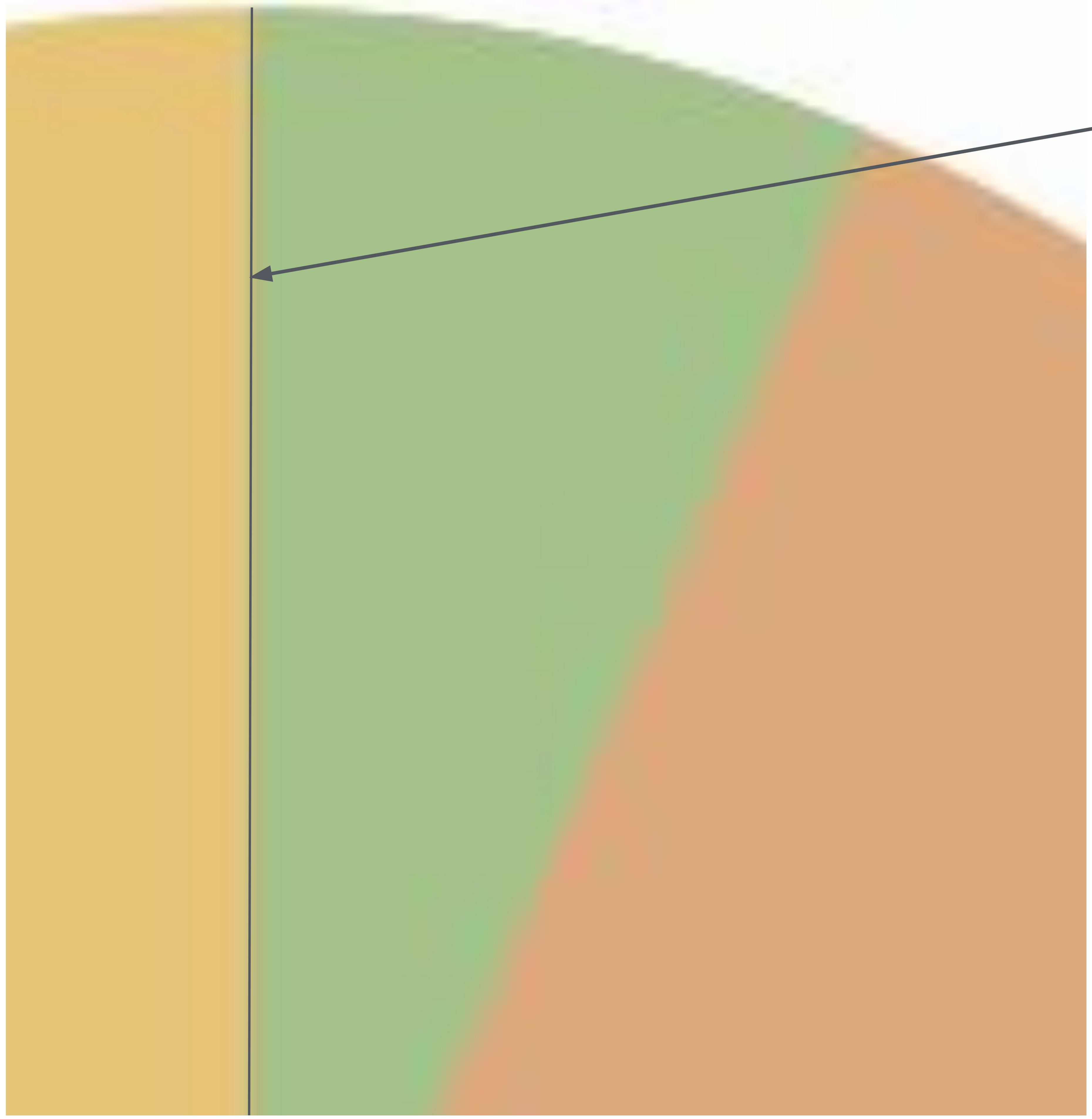
3593 **stable** languages:
Not sustained by formal institutions; norm at home for children to learn and use the language

3072 **endangered** languages:
No longer the norm for children to learn and use the language

451 **extinct** languages:
Languages no longer used; no one identifies with the language

Image derived from the language vitality data from Ethnologue's website ("Eberhard, et al., 2024. Ethnologue: Languages of the World. Twenty-seventh edition. Dallas, Texas: SIL International.). Image approx. at scale.

Languages by Vitality

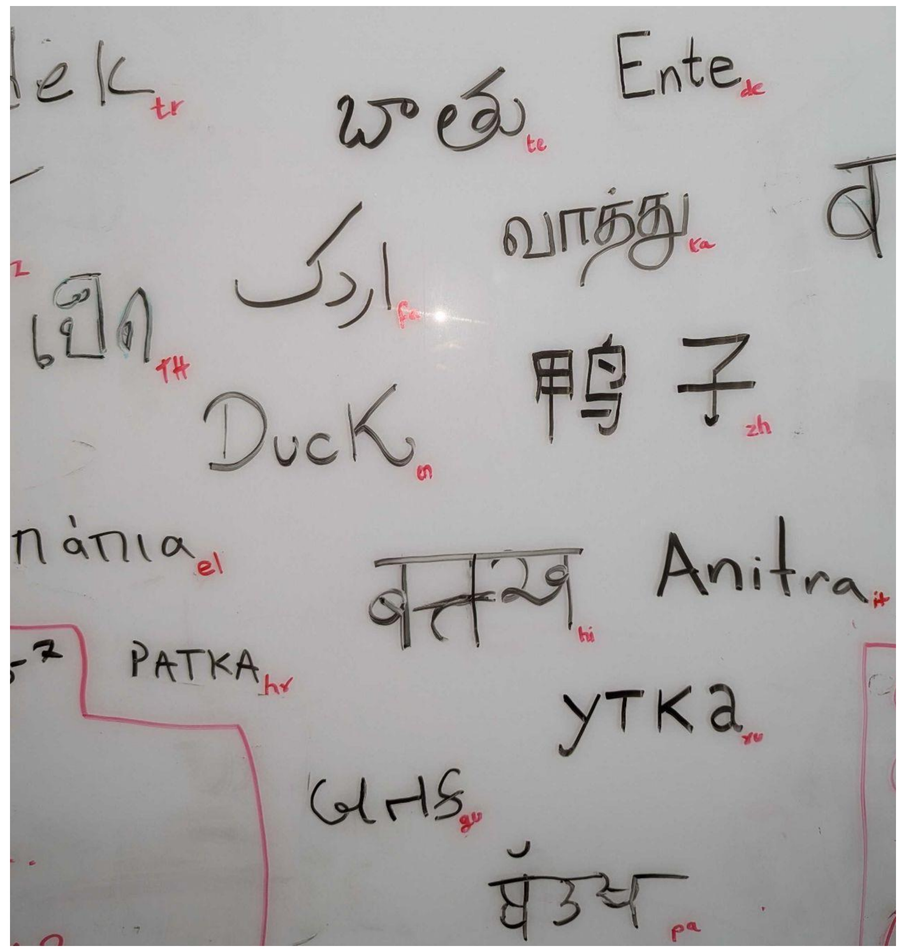


That thin black line is **English**!

Image derived from the language vitality data from Ethnologue's website ("Eberhard, et al., 2024. Ethnologue: Languages of the World. Twenty-seventh edition. Dallas, Texas: SIL International.). Image approx. at scale.

How do Languages Differ?

Scripts:



Word Order:

English: I met Jack. (SVO order)
Hindi : मैं जैक से मिला। (SOV order)
Filipino: Nakilala ko si Jack. (VSO order)

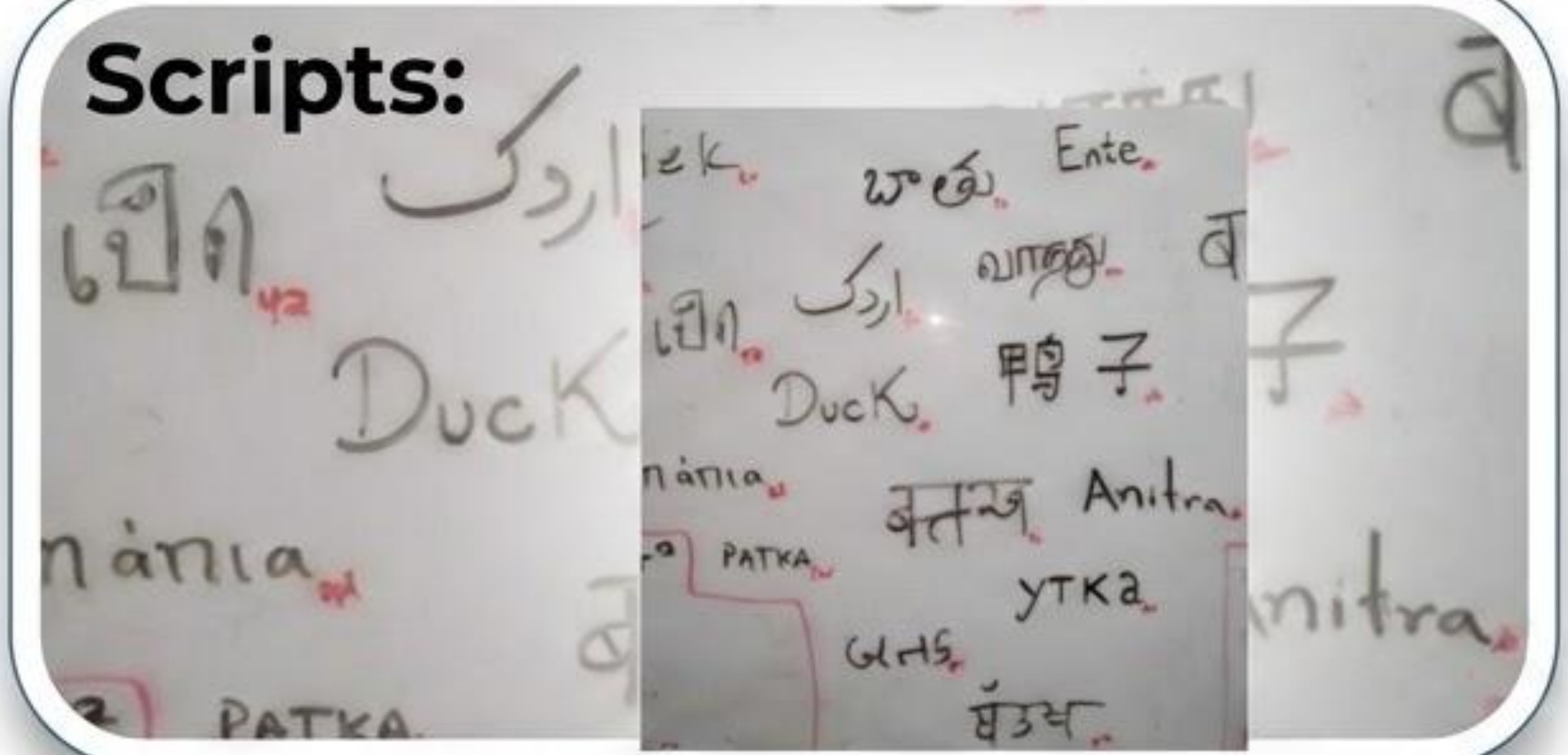
Semantic Variations:

dara : door (*Farsi*) vs burrow (*Gujarati*)
śikśā: education (*Hindi*) vs punishment (*Gujarati*)

And many more ……

How do Languages Differ?

Scripts:



Word Order:

English: I met Jack. → (SVO order)
Hindi → मैं जैक से मिला। → (SOV order)
Filipino: Nakilala ko si Jack. (VSO order)

Semantic Variations:
dara : door (*Farsi*) ↔ burrow (*Gujarati*)
śikśā: education (*Hindi*) ↔ punishment (*Gujarati*)

And many more ......

Multilingual LLMs

How can we train Llama 5 200B on 100 Languages?

Moving part?

- Data
  - Source?
  - Cleaning?
  - Balancing?
- Tokenizer:
  - Size, data mixture
- Architecture Design:
  - Dense or MoE?
  - Hidden dimension?
  - Context window?

Multilingual LLMs: Overview

- LLMs that support multiple languages
  - Parameters shared across languages
  - Trained on a large amount of multilingual data (unlabeled & labeled)
  - Often rely on cross-lingual transfer abilities across languages

Incidentally Multilingual Models

Mistral 7B

Gemma

Claude 3

Llama 2

Meta AI

Multilingual LLMs: Overview

- LLMs that support multiple languages
  - Parameters shared across languages
  - Trained on a large amount of multilingual data (unlabeled & labeled)
  - Often rely on cross-lingual transfer abilities across languages

Incidentally Multilingual Models

Mistral 7B

Gemma

Claude 3

Llama 2

Meta AI

Natively Multilingual Models

mT5

BLOOM

Okapi

MALA-500

Aya Model

Aya 23

Closed Data Models

Llama 3.1

Gemini

Phi-3

GPT - 4

Moving part?

Data

- Source?
- Cleaning?
- Balancing?

Tokenizer

- Size, data mixture

Architecture Design

- Dense or MoE?
- Hidden dimension?
- Context window?

Directions of Innovations in Multilingual LLMs

Data

Infrastructure

Directions of Innovations in Multilingual LLMs

Data

- Methods to efficiently procure labeled & unlabeled data
  - Quality vs Quantity trade-off
  - Impact of data diversity
- Alignment data collection strategies

Infrastructure

Directions of Innovations in Multilingual LLMs

Data

- Methods to efficiently procure labeled & unlabeled data
  - Quality vs Quantity trade-off
  - Impact of data diversity
- Alignment data collection strategies

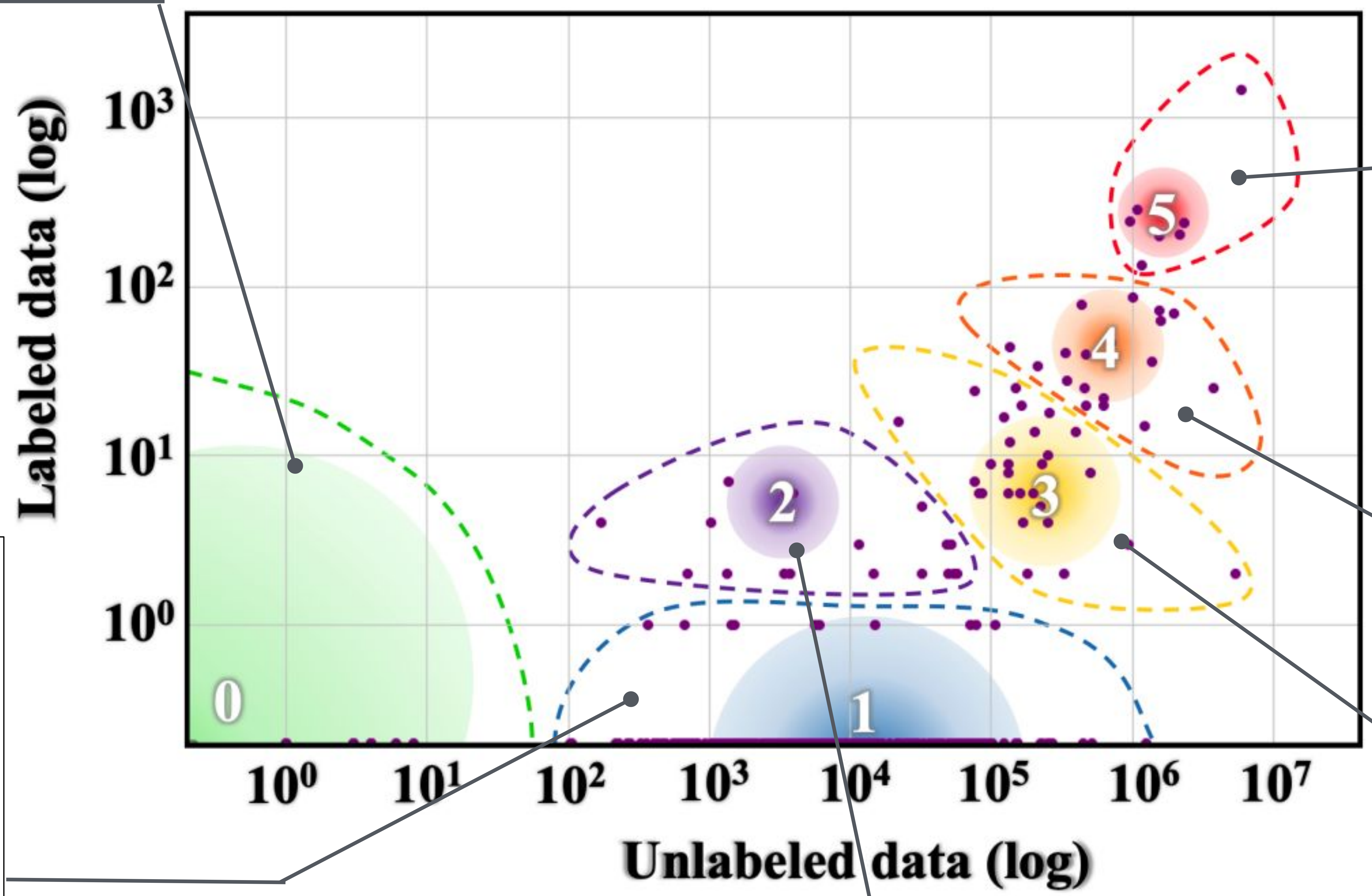
Infrastructure

- Breaking the curse of multilinguality (more on this if time permits)
- Extending LLMs to unseen languages
- Efficient tokenization for low-resource languages

We'll focus on the **Data** direction today!

Languages of the World via the Data Lens

"The Left-Behinds"
Impossible effort required to lift them into digital space
#Langs: 2191
E.g.: Warlpiri, Gaelic, Gondi
#Speakers: 1.2B



"The Winners"
the quintessential rich-resource languages
#Langs: 7
E.g.: English, German, French
#Speakers: 2.5B

"The Underdogs"
dedicated NLP communities conducting research on these languages
#Langs: 18
E.g.: Russian, Dutch, Korean
#Speakers: 2.2B

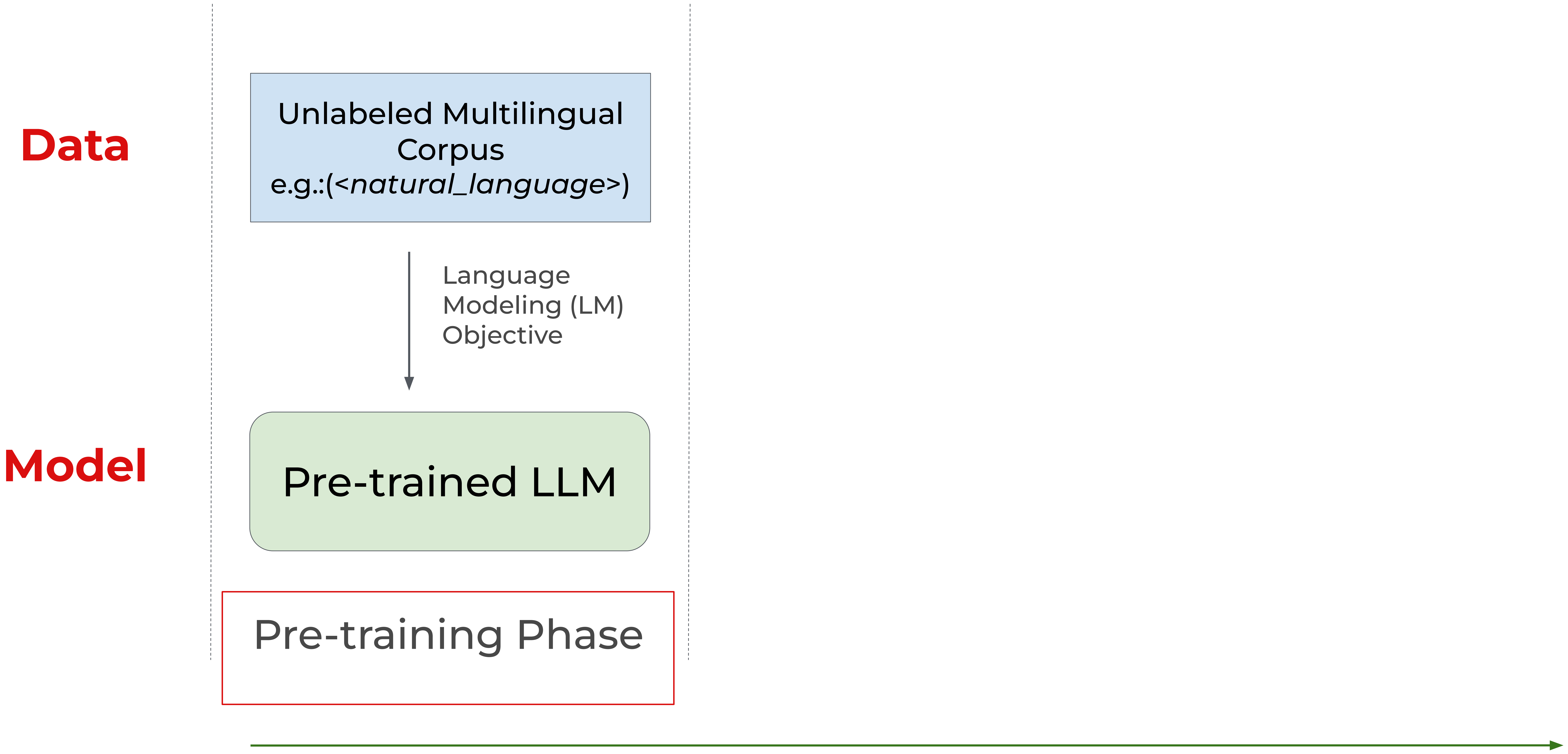
"The Scraping-Bys"
Need solid, organized movement that increases awareness
#Langs: 222
E.g.: Nepali, Gujarati, Armenian
#Speakers: 30M

"The Rising Stars"
let down by insufficient efforts in labeled data collection
#Langs: 28
E.g.: Hebrew, Ukrainian, Urdu
#Speakers: 1.8B

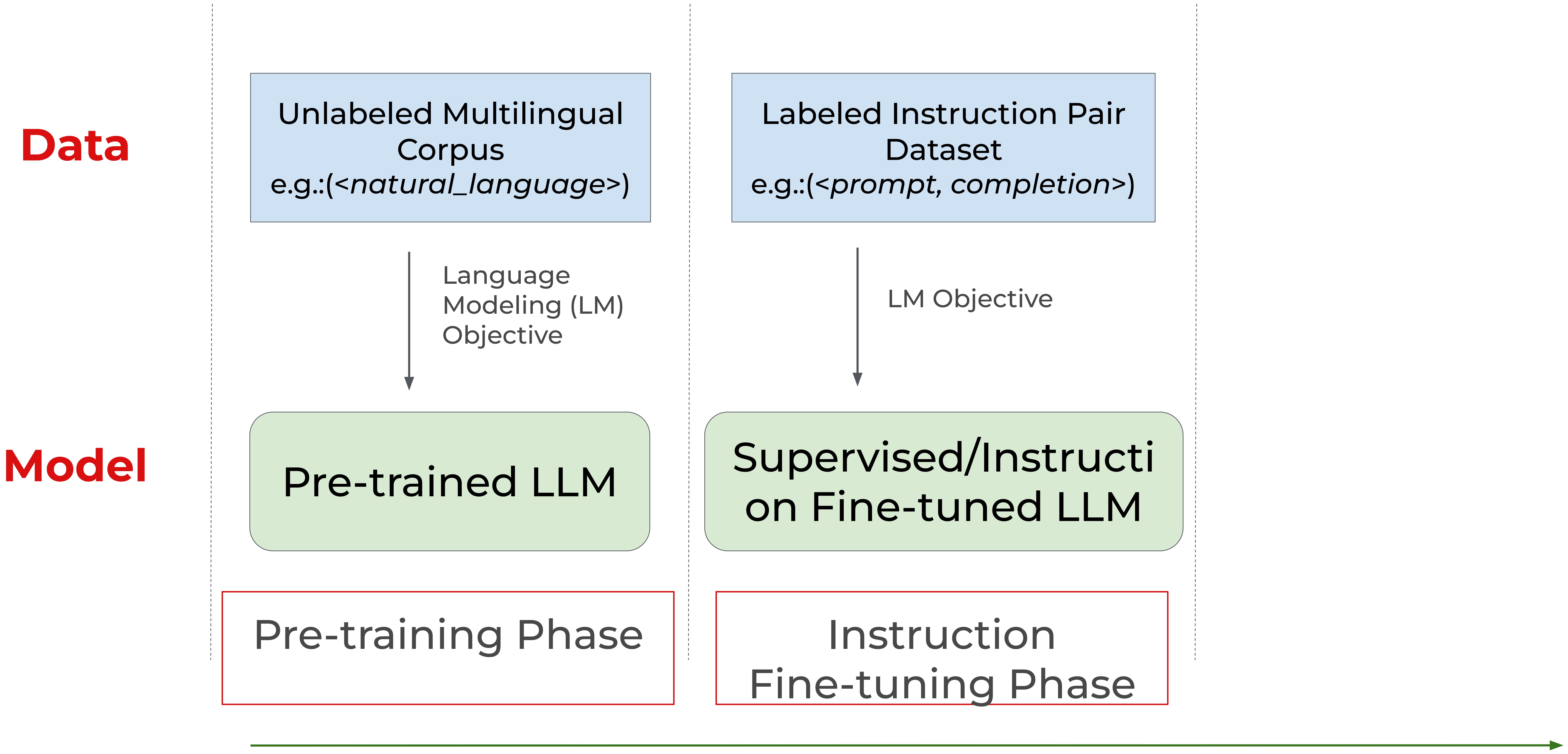
"The Hopefuls"
languages still fight on with their gasping breath
#Langs: 19 ; E.g.: Marathi, Irish, Yoruba
#Speakers: 5.7M

Figure from [The State and Fate of Linguistic Diversity and Inclusion in the NLP World](#) (Joshi et al., ACL 2020)
For language categorization of your language see:
https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt

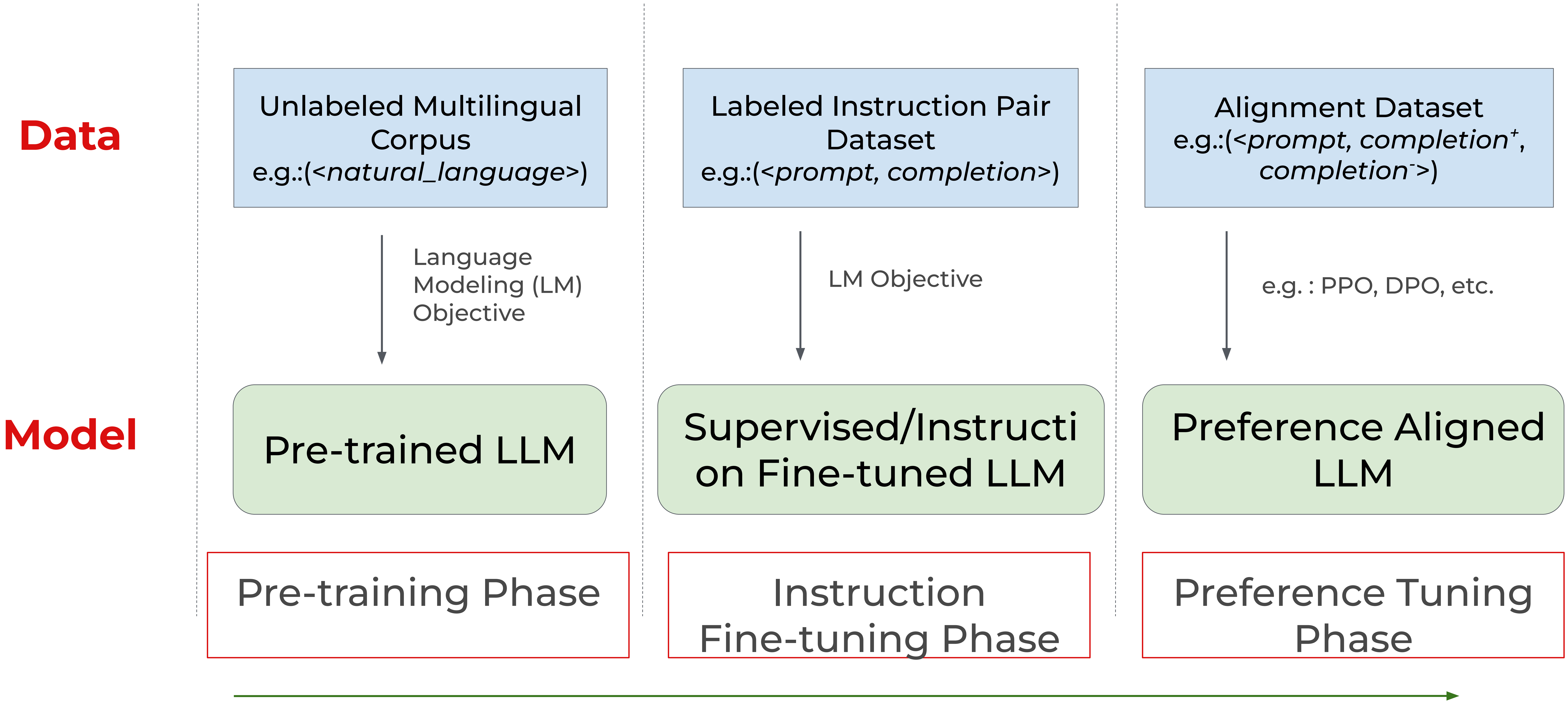
The Multilingual LLM Pipeline



The Multilingual LLM Pipeline



The Multilingual LLM Pipeline



Multilingual Pre-training

Multilingual Pre-training: mC4

- **Multilingual C4 (mC4)[1] [6.6B pages, 6.3T tokens]**
  - **C4**: Colossal Clean Crawled Corpus[2]

[1] mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)
[2] Exploring the limits of transfer learning with a unified text-to-text transformer. (Raffel et al.,, JMLR 2020)
* - https://pypi.org/project/langdetect/ (Only pages with a probability 99% or higher of being English were considered)
$ - https://github.com/google/cld3 (Pages with a language confidence of below 70% were discarded)

Multilingual Pre-training: mC4

- **Multilingual C4 (mC4)[1] [6.6B pages, 6.3T tokens]**
  - **C4**: Colossal Clean Crawled Corpus[2]
    - Cleaned version of the Common Crawl's snapshot of the internet (April 2019)
    - Filtered for pages predominantly English as per a language detector

[1] mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)
[2] Exploring the limits of transfer learning with a unified text-to-text transformer. (Raffel et al.,, JMLR 2020)
* - https://pypi.org/project/langdetect/ (Only pages with a probability 99% or higher of being English were considered)
$ - https://github.com/google/cld3 (Pages with a language confidence of below 70% were discarded)

Multilingual Pre-training: mC4

- **Multilingual C4 (mC4)[1] [6.6B pages, 6.3T tokens]**
  - **C4**: Colossal Clean Crawled Corpus[2]
    - Cleaned version of the Common Crawl's snapshot of the internet (April 2019)
    - Filtered for pages predominantly English as per a language detector
  - Use 71 snapshots of Common Crawl
  - Supports **101 languages** (with 6 languages in two scripts)
    - Identified using the *cld3* language detector
  - Other filters: length, deduplication, profanity, etc.

[1] mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)
[2] Exploring the limits of transfer learning with a unified text-to-text transformer. (Raffel et al.,, JMLR 2020)
* - https://pypi.org/project/langdetect/ (Only pages with a probability 99% or higher of being English were considered)
$ - https://github.com/google/cld3 (Pages with a language confidence of below 70% were discarded)

Multilingual Pre-training: mC4

- **Multilingual C4 (mC4)[1] [6.6B pages, 6.3T tokens]**
  - **C4**: Colossal Clean Crawled Corpus[2]
    - Cleaned version of the Common Crawl's snapshot of the internet (April 2019)
    - Filtered for pages predominantly English as per a language detector
  - Use 71 snapshots of Common Crawl
  - Supports **101 languages** (with 6 languages in two scripts)
    - Identified using the *cld3* language detector
  - Other filters: length, deduplication, profanity, etc.
- Models trained on mC4: mT5, mT0, Aya-101

[1] mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)
[2] Exploring the limits of transfer learning with a unified text-to-text transformer. (Raffel et al.,, JMLR 2020)
* - https://pypi.org/project/langdetect/ (Only pages with a probability 99% or higher of being English were considered)
$ - https://github.com/google/cld3 (Pages with a language confidence of below 70% were discarded)

Multilingual Pre-training: mC4

- **Multilingual C4 (mC4)[1] [6.6B pages, 6.3T tokens]**
  - **C4**: Colossal Clean Crawled Corpus[2]
    - Cleaned version of the Common Crawl's snapshot of the internet

Figure 1: Page counts per language in mC4 (left axis) from mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)

[1] Exploring the limits of transfer learning with a unified text-to-text transformer. (Raffel et al.,, JMLR 2020)
[2] mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)
* - https://pypi.org/project/langdetect/ (Only pages with a probability 99% or higher of being English were considered)
$ - https://github.com/google/cld3 (Pages with a language confidence of below 70% were discarded)

Multilingual Pre-training: mC4

- **Multilingual C4 (mC4)[1] [6.6B pages, 6.3T tokens]**
    - **C4**: Colossal Clean Crawled Corpus[2]
        - Cleaned version of the Common Crawl's snapshot of the internet

What could go wrong here?

Figure 1: Page counts per language in mC4 (left axis) from [mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer](#) (Xue et al., NAACL 2021)

[1] Exploring the limits of transfer learning with a unified text-to-text transformer. (Raffel et al.,, JMLR 2020)
[2] mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)
* - https://pypi.org/project/langdetect/ (Only pages with a probability 99% or higher of being English were considered)
$ - https://github.com/google/cld3 (Pages with a language confidence of below 70% were discarded)

Multilingual Pre-training: mC4

- **Multilingual C4 (mC4)[1] [6.6B pages, 6.3T tokens]**
  - **C4**: Colossal Clean Crawled Corpus[2]
    - Cleaned version of the Common Crawl's snapshot of the internet

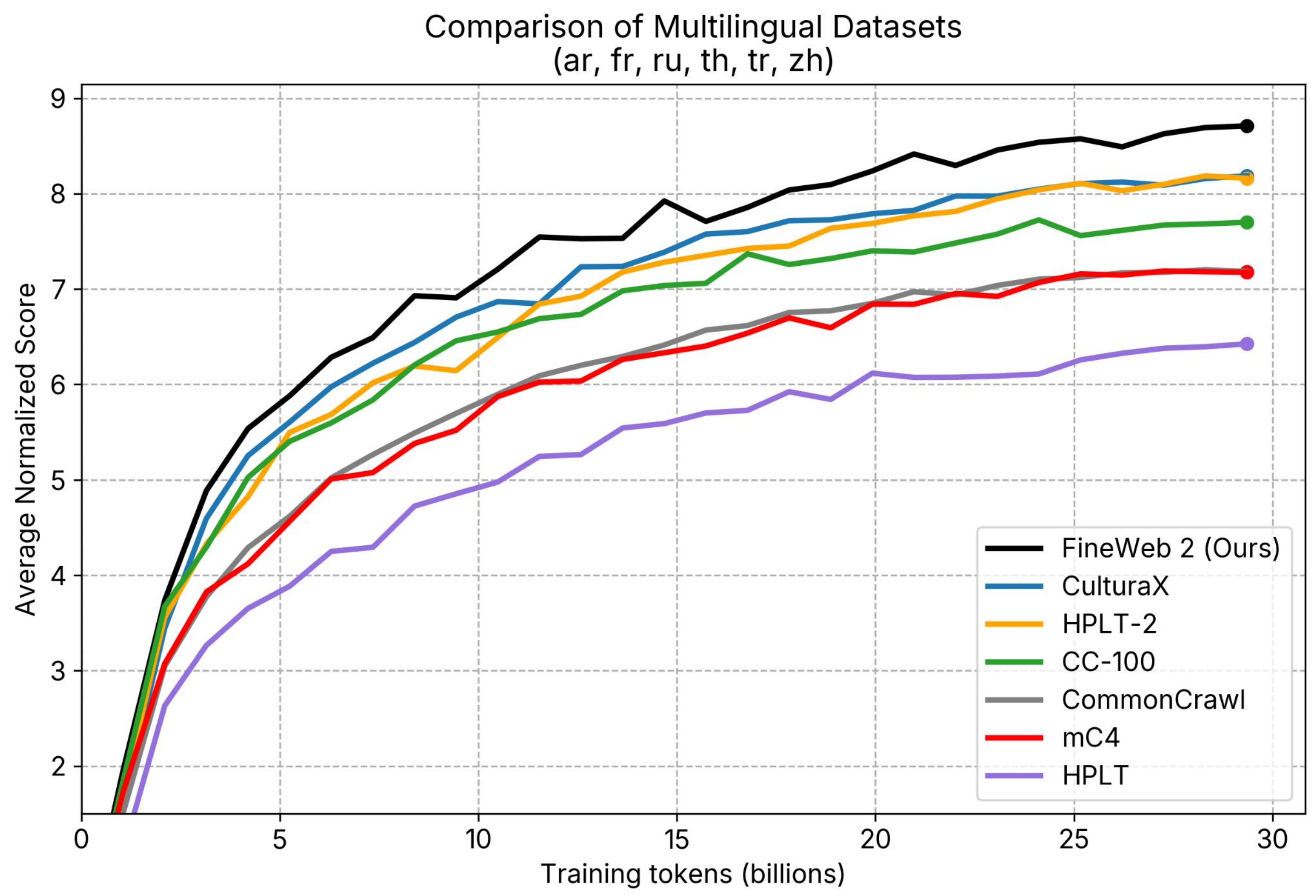
Figure 1: Page counts per language in mC4 (left axis) from mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)

Sampling affects model performance*:
- If low-resource languages are highly sampled too often, the model may overfit
- If high-resource languages are not trained on enough, the model will underfit

[1] Explori
[2] mT5: A
* - https:/
$ - https:/

Multilingual Pre-training: Glot500-c

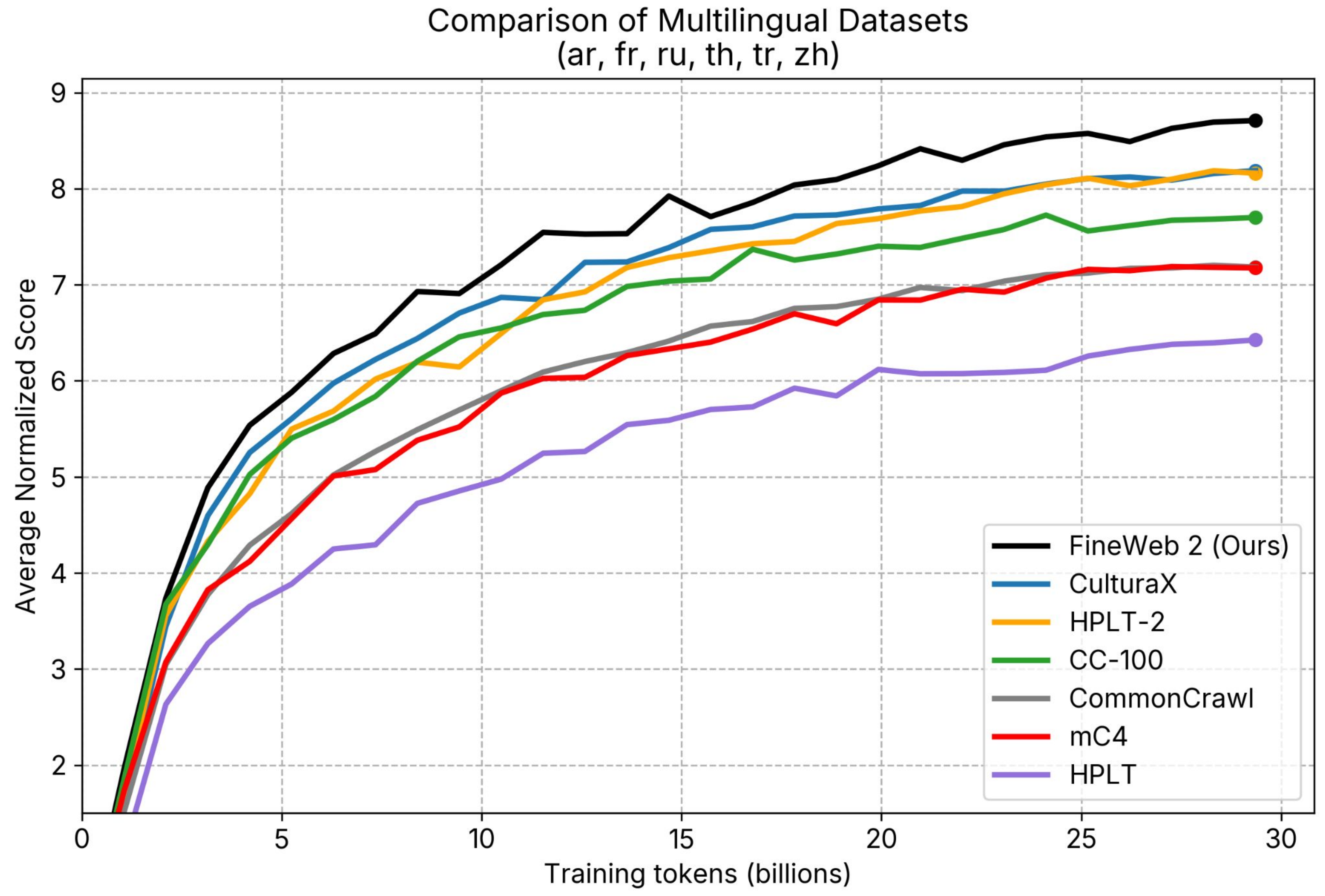
- **Fineweb-2[1] [1.5B sentences, 20 TB]**
  - 96 CommonCrawl snapshots on 1,868 languages.
  - 5 billion documents, with over 3 trillion words



[1] FineWeb2 Dataset
[2] FineWeb2: One Pipeline to Scale Them All -- Adapting Pre-Training Data Processing to Every Language

Multilingual Pre-training: Glot500-c

- **Fineweb-2[1] [1.5B sentences, 20 TB]**
  - 96 CommonCrawl snapshots on 1,868 languages.
  - 5 billion documents, with over 3 trillion words
    - Language Identification and filtering 🔍
    - Deduplication per language 🔄
    - Filtering per language 🧹
    - PII Anonymization and fixes 🎭



[1] FineWeb2 Dataset
[2] FineWeb2: One Pipeline to Scale Them All -- Adapting Pre-Training Data Processing to Every Language

Multilingual Pre-training: Glot500-c

- **Glot500-c[1] [1.5B sentences, 600 GB]**
  - Subset of **Glot2000-c** that covers 2266 languages:
    - Diverse data sources: religious texts, news articles, scientific papers, etc.
    - Several filters:
      - Chunk-level filters$
      - Corpus-level filters
  - Set of **511 languages*** with > 30k chunks

[1] Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages (Imani et al., ACL 2023)
* - They cover 30 scripts. They also count a distinct language-script pair as a separate pair
$ - The chunk-level filters are taken from BigScience's ROOTS Corpus (The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset(Laurençon et al., NeurIPS 2022)). This was used to train models like BLOOM, BLOOMZ, etc.

Multilingual Pre-training: Glot500-c

- **Glot500-c[1] [1.5B sentences, 600 GB]**
  - Subset of **Glot2000-c** that covers 2266 languages:
    - Diverse data sources: religious texts, news articles, scientific papers, etc.
    - Several filters:
      - Chunk-level filters$
      - Corpus-level filters
  - Set of <u>**511 languages**</u>* with > 30k chunks

SF1 Character repetition. If the ratio of repeated characters is too high, it is likely that the sentence has not enough textual content.

SF2 Word repetition. A high ratio of repeated words indicates non-useful repetitive content.

SF3 Special characters. Sentences with a high ratio of special characters are likely to be crawling artifacts or computer code.

SF4 Insufficient number of words. Since training language models requires enough context, very small chunks of text are not useful.

SF5 Deduplication. If two sentences are identical after eliminating punctuation and white space, one is removed.

[1] Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages (Imani et al., ACL 2023)
* - They cover 30 scripts. They also count a distinct language-script pair as a separate pair
$ - The chunk-level filters are taken from BigScience's ROOTS Corpus (The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset(Laurençon et al., NeurIPS 2022)). This was used to train models like BLOOM, BLOOMZ, etc.

Multilingual Pre-training: Glot500-c

- **Glot500-c[1] [1.5B sentences, 600 GB]**
  - Subset of **Glot2000-c** that covers 2266 languages:
    - Diverse data sources: religious texts, news articles, scientific papers, etc.
    - Several filters:
      - Chunk-level filters
      - Corpus-level filters
  - Set of **<u>511 languages</u>*** with > 30k chunks

Corpus-level filters detect if the corpus of a language-script is noisy; e.g., the corpus is in another language or consists of non-meaningful content such as tabular data. We employ filters CF1 and CF2.

CF1 In case of **mismatch between language and script**, the corpus is removed; e.g., Chinese written in Arabic is unlikely to be Chinese.

CF2 Perplexity mismatch. For each language-script L1, we find its closest language-script L2: the language-script with the lowest perplexity divergence (§3.3). If L1 and L2 are not in the same typological family, we check L1/L2 manually and take appropriate action such as removing the corpus (e.g., if it is actually English) or correcting the ISO code assigned to the corpus.

[1] Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages (Imani et al., ACL 2023)
* - They cover 30 scripts. They also count a distinct language-script pair as a separate pair

Multilingual Pre-training: Glot500-c

- **Glot500-c[1] [1.5B sentences, 600 GB]**
  - Subset of **Glot2000-c** that covers 2266 languages:
    - Diverse data sources: religious texts, news articles, scientific papers, etc.
    - Several filters:
      - Chunk-level filters
      - Corpus-level filters
  - Set of **511 languages*** with > 30k chunks
- Models trained on Glot500-c: Glot500-m, MALA-500

Corpus-level filters detect if the corpus of a language-script is noisy; e.g., the corpus is in another language or consists of non-meaningful content such as tabular data. We employ filters CF1 and CF2.

CF1 In case of **mismatch between language and script**, the corpus is removed; e.g., Chinese written in Arabic is unlikely to be Chinese.

CF2 Perplexity mismatch. For each language-script L1, we find its closest language-script L2: the language-script with the lowest perplexity divergence (§3.3). If L1 and L2 are not in the same typological family, we check L1/L2 manually and take appropriate action such as removing the corpus (e.g., if it is actually English) or correcting the ISO code assigned to the corpus.

[1] Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages (Imani et al., ACL 2023)
* - They cover 30 scripts. They also count a distinct language-script pair as a separate pair

Multilingual Instruction Fine-tuning

Moving part?

Data

- Source?
- Cleaning?
- Balancing?

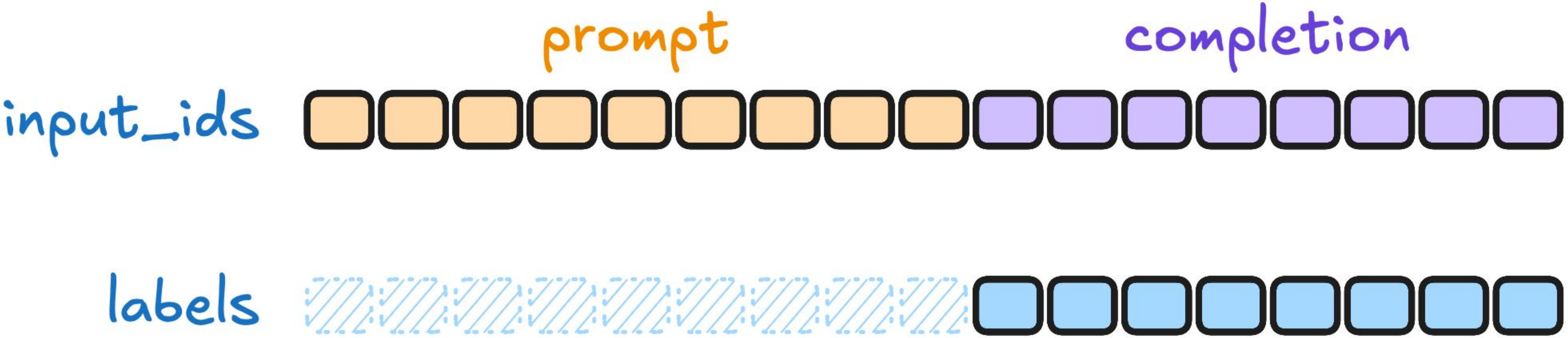
Tokenizer

- Size, data mixture

Architecture Design

- Dense or MoE?
- Hidden dimension?
- Context window?

Template-based



Template-based

Input	Output
Jim, I had a lot of fun at dinner …	Not spam
Congratulations! You just won …	Spam
…	….

Template-based

Input	Output
Jim, I had a lot of fun at dinner …	Not spam
Congratulations! You just won …	Spam
…	….

Instruction template

Prompt	Completion
Jim, I had a lot of fun at dinner … Indicate if this mail is spam or not. This mail is …	not a spam
Congratulations! You just won …Indicate if this mail is spam or not. This mail is …	spam
…	….

Template-based

- Convert existing multilingual datasets to prompt-completion pairs
- Instructions can be English or multilingual
- Easy to scale

[1] Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks (Wang et al., EMNLP 2022) (largely collected via class-sourcing and public invitation)
[2] Crosslingual generalization through multitask finetuning (Muennighoff et al., ACL 2023) (xP3mt translated using Google Translate API)

Template-based

- Convert existing multilingual datasets to prompt-completion pairs
- Instructions can be English or multilingual
- Easy to scale
- Drawback?

[1] Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks (Wang et al., EMNLP 2022) (largely collected via class-sourcing and public invitation)
[2] Crosslingual generalization through multitask finetuning (Muennighoff et al., ACL 2023) (xP3mt translated using Google Translate API)

Template-based

- Convert existing multilingual datasets to prompt-completion pairs
- Instructions can be English or multilingual
- Easy to scale
- Low in diversity

[1] Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks (Wang et al., EMNLP 2022) (largely collected via class-sourcing and public invitation)
[2] Crosslingual generalization through multitask finetuning (Muennighoff et al., ACL 2023) (xP3mt translated using Google Translate API)

Template-based

- Convert existing multilingual datasets to prompt-completion pairs
- Instructions can be English or multilingual
- Easy to scale
- Low in diversity
- Datasets:
  - **Supernatural Instructions[1]:** 76 task types, 55 languages, English instructions
  - **xP3 and xP3mt[2]:** 16 task types, 46 languages
    - **xP3** has English instructions while **xP3mt** is its machine-translated version

[1] Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks (Wang et al., EMNLP 2022) (largely collected via class-sourcing and public invitation)
[2] Crosslingual generalization through multitask finetuning (Muennighoff et al., ACL 2023) (xP3mt translated using Google Translate API)

Template-based

- Convert existing multilingual datasets to prompt-completion pairs
- Instructions can be English or multilingual
- Easy to scale
- Low in diversity
- Datasets:
    - **Supernatural Instructions[1]:** 55 languages, 76 task types, English instructions
    - **xP3 and xP3mt[2]:** 46 languages, 13 task types
        - **xP3** has English instructions while **xP3mt** is its machine-translated version
    - **xP3x[3]:** xP3 extended to 277 languages, 16 task types
        - Pruned through a human-auditing process
    - **Aya Collection[4]:** 74 languages, 14 task types, Human-written multilingual instructions

    and more …

[1] Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks (Wang et al., EMNLP 2022) (largely collected via class-sourcing and public invitation)
[2] Crosslingual generalization through multitask finetuning (Muennighoff et al., ACL 2023) (xP3mt translated using Google Translate API)
[3] Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., ACL 2024)
[4] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)

Translation-based

- Templates lack diversity
- Translate diverse English instructions into other languages
  - Popular machine translation models[1,2] to the rescue!

[1] Google Translate API
[2] No language left behind: Scaling human-centered machine translation (NLLB-Team.,2022)
[3] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)
[4] Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., ACL 2024)

Do Translated Instructions over English Ones Help?

Unseen Tasks

Trained on xP3 (English-only)

Trained on xP3mt

Task	Prompt	Average accuracy			
		BLOOMZ	BLOOMZ-MT	mT0-13B	mT0-13B-MT
XNLI	EN	**52.99**	49.01	48.24	**51.29**
	MT	37.56	**41.16**	39.31	**41.66**
	HT	40.4	**43.88**	44.95	**46.87**
XCOPA	EN	72.52	**73.24**	**81.4**	80.36
	MT	70.04	**71.84**	**81.16**	79.64
XStoryCloze	EN	**81.73**	81.39	81.99	**82.3**
	MT	80.89	**81.76**	**83.37**	82.86
XWinograd	EN	**60.07**	59.15	70.49	**73.24**
	MT	58.48	**60.14**	66.89	**72.33**

Table 1: Comparison between EN (English), MT (machine-translated) and HT (human-translated) prompts for 176B BLOOMZ and 13B mT0 models finetuned on either only English or English and machine-translated multilingual prompts (-MT).

Table from Crosslingual generalization through multitask finetuning (Muennighoff et al., ACL 2023)

Translated instructions usually result in improved performance

Translation-based

- Templates lack diversity
- Translate diverse English instructions into other languages
  - Popular machine translation models[1,2] to the rescue!
- Bottleneck?

[1] Google Translate API
[2] No language left behind: Scaling human-centered machine translation (NLLB-Team.,2022)
[3] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)
[4] Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., ACL 2024)

Translation-based

- Templates lack diversity
- Translate diverse English instructions into other languages
  - Popular machine translation models[1,2] to the rescue!
- Bottleneck?
  - Translation quality in lower resourced languages

[1] Google Translate API
[2] No language left behind: Scaling human-centered machine translation (NLLB-Team.,2022)
[3] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)
[4] Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., ACL 2024)

Translation-based

- Templates lack diversity
- Translate diverse English instructions into other languages
  - Popular machine translation models[1,2] to the rescue!
- Bottleneck?
  - Translation quality in lower resourced languages
  - Introduction of translation artefacts known as *translationese*

[1] Google Translate API
[2] No language left behind: Scaling human-centered machine translation (NLLB-Team.,2022)
[3] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)
[4] Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., ACL 2024)

Translation-based

- Templates lack diversity
- Translate diverse English instructions into other languages
  - Popular machine translation models[1,2] to the rescue!
- Bottleneck?
  - Translation quality in lower resourced languages
  - Introduction of translation artefacts known as *translationese*
- Datasets:
  - **Aya Collection[3]:** 101 languages, 19 datasets
    - Diverse sources: xP3, Flan Collection, Dolly, etc.; Translated using NLLB[1]
  - **ShareGPT-Command[4]:** 93 languages
    - ShareGPT: Synthetic English completions from Command for human prompts
    - Translate prompt-completion pairs using NLLB

[1] Google Translate API
[2] No language left behind: Scaling human-centered machine translation (NLLB-Team.,2022)
[3] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)
[4] Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., ACL 2024)

Human Annotations

- **Gold** standard

[1] Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences (Pinhanez et al., 2024)
[2] A Survey of Corpora for Germanic Low-Resource Languages and Dialects (Blaschke et al., NoDaLiDa 2023)
[3] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)

Human Annotations

- **Gold** standard
- Expensive to collect
  - **Technological factors:** Support of languages on annotation platforms
  - **Sociological factors:**
    - Access to language technology[1]
    - Dialectical and other biases[2]

[1] Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences (Pinhanez et al., 2024)
[2] A Survey of Corpora for Germanic Low-Resource Languages and Dialects (Blaschke et al., NoDaLiDa 2023)
[3] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)

Human Annotations

- **Gold** standard
- Expensive to collect
    - **Technological factors:** Support of languages on annotation platforms
    - **Sociological factors:**
        - Access to language technology[1]
        - Dialectical and other biases[2]
- Dataset:
    - **Aya Dataset[3]:** 65 languages, 2k contributors across 110 countries
        - Created a multi-platform Annotation platform - **Aya Annotation Platform**
        - Instances human annotated, re-annotated & feedback curated
        - Implement leaderboarding via **Aya Score** to boost quality

[1] Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences (Pinhanez et al., 2024)
[2] A Survey of Corpora for Germanic Low-Resource Languages and Dialects (Blaschke et al., NoDaLiDa 2023)
[3] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)

Which Approach is the Best?

% of the training budget

Weighting name	Human Annot.: Aya Dataset	Template: Aya Templates	Template: xP3x	Template: Data Provenance	Translation: Aya Translations	Translation: ShareGPT-Command
Human Annot. Heavy	25	4	20	6	30	15
Translation Heavy	10	1.5	15	3.5	47.5	22.5
Template Heavy	20	10	30	10	20	10

Model	Base Model	IFT Mixture	Held out tasks (Accuracy %) XCOPA	XNLI	XSC	XWG	Avg
46 Languages							
mT0	mT5 13B	xP3	75.6	55.3	87.2	73.6	72.9
BLOOMZ	BLOOM 176B	xP3	64.3	52.0	82.6	63.3	65.5
52 Languages							
Bactrian-X 13B	Llama 13B	Bactrian-X	52.4	34.5	51.8	50.5	47.3
101 Languages							
mT0x	mT5 13B	xP3x	71.7	45.9	85.1	60.6	65.8
Aya (human-anno-heavy)	mT5 13B	All Mixture	76.5	**59.2**	89.3	70.6	73.9
Aya (template-heavy)	mT5 13B	All Mixture	**77.3**	58.3	**91.2**	**73.7**	**75.1**
★**Aya** (translation-heavy)	mT5 13B	All Mixture	76.7	58.3	90.0	70.7	73.9

Table 5: Results for held-out task evaluation. Results are averaged across all splits of XCOPA, XNLI, XStoryCloze, and XWinoGrad. ★**Aya** (translation-heavy) is used as the final **Aya** model. See § 5.6 for detailed analysis.

- Aya-101 outperforms all other contemporary models (even BLOOMZ 176B)
- Template-heavy seems to be the best fine-tuning mixture

Tables from Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., ACL 2024)

Which Approach is the Best?

Weighting name	HUMAN ANNOT. Aya Dataset	TEMPLATE Aya Templates	xP3x	Data Provenance	TRANSLATION Aya Translations	ShareGPT-Command
Human Annot. Heavy	25	4	20	6	30	15
Translation Heavy	10	1.5	15	3.5	47.5	22.5
Template Heavy	20	10	30	10	20	10

Model	Base Model	IFT Mixture	Held out tasks (Accuracy %) XCOPA	XNLI	XSC	XWG	Avg
46 LANGUAGES							
MT0	mT5 13B	xP3	75.6	55.3	87.2	73.6	72.9
BLOOMZ	BLOOM 176B	xP3	64.3	52.0	82.6	63.3	65.5
52 LANGUAGES							
BACTRIAN-X 13B	Llama 13B	Bactrian-X	52.4	34.5	51.8	50.5	47.3
101 LANGUAGES							
MT0X	mT5 13B	xP3x	71.7	45.9	85.1	60.6	65.8
Aya (human-anno-heavy)	mT5 13B	All Mixture	76.5	**59.2**	89.3	70.6	73.9
Aya (template-heavy)	mT5 13B	All Mixture	**77.3**	58.3	**91.2**	**73.7**	**75.1**
★**Aya** (translation-heavy)	mT5 13B	All Mixture	76.7	58.3	90.0	70.7	73.9

Table 5: Results for held-out task evaluation. Results are averaged across all splits of XCOPA, XNLI, XStoryCloze, and XWinoGrad. ★**Aya** (translation-heavy) is used as the final **Aya** model. See § 5.6 for detailed analysis.

Model	IFT Mixture	Generative Tasks FLORES-200 (spBleu) X→ En	En → X	XLSum (RougeLsum)	Tydi-QA (F1)
101 LANGUAGES					
MT0X	xP3x	20.2	14.5	21.4	76.1
Aya (human-anno-heavy)	All Mixture	25.1	18.9	22.2	77.9
Aya (templated-heavy)	All Mixture	25.0	18.6	**23.2**	**78.8**
★**Aya** (translation-heavy)	All Mixture	**29.1**	**19.0**	22.0	77.8

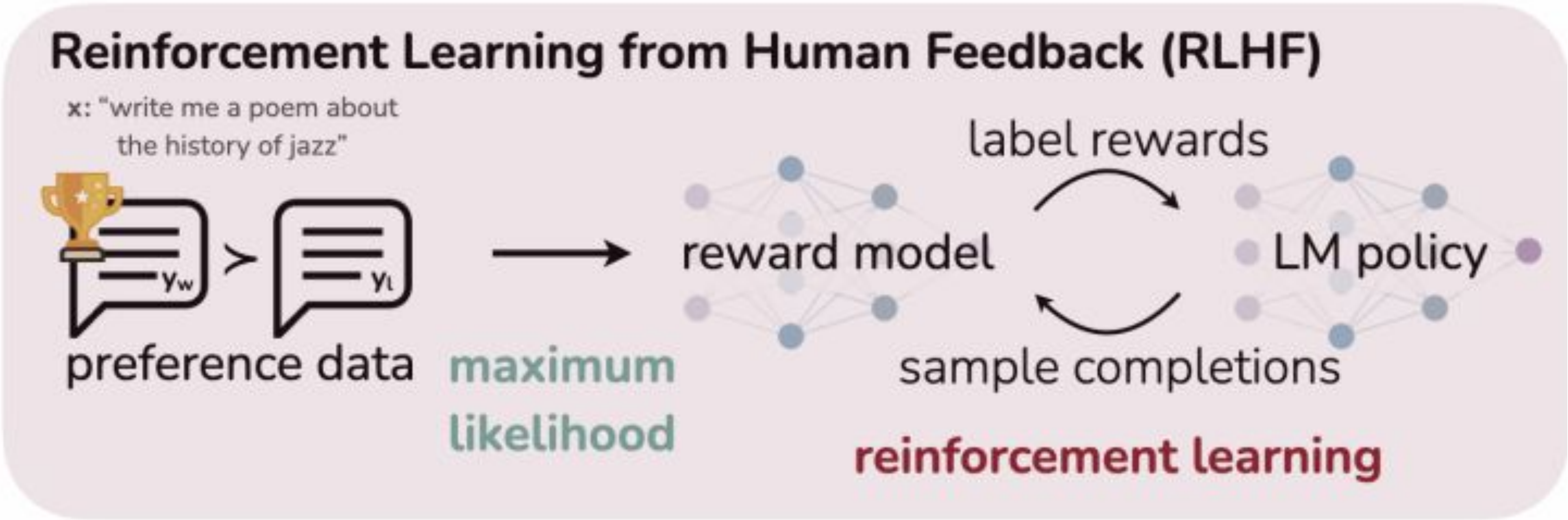
Table 7: Generative tasks' results for mT0x and **Aya** model variants based on different weighting ablations. Here the **translation-heavy** weighting has the highest spBleu score on Flores and the **template-heavy** weighting has the highest RougeLsum and F1 scores on XLSum and Tydiqa respectively. ★**Aya** (translation-heavy) is used as the final **Aya** model. See § 5.6 for detailed analysis.

Translation-heavy performs better on translation tasks; template-heavy is better on other generative tasks

Tables from Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., ACL 2024)

Multilingual Alignment

Online vs Offline Alignment Methods

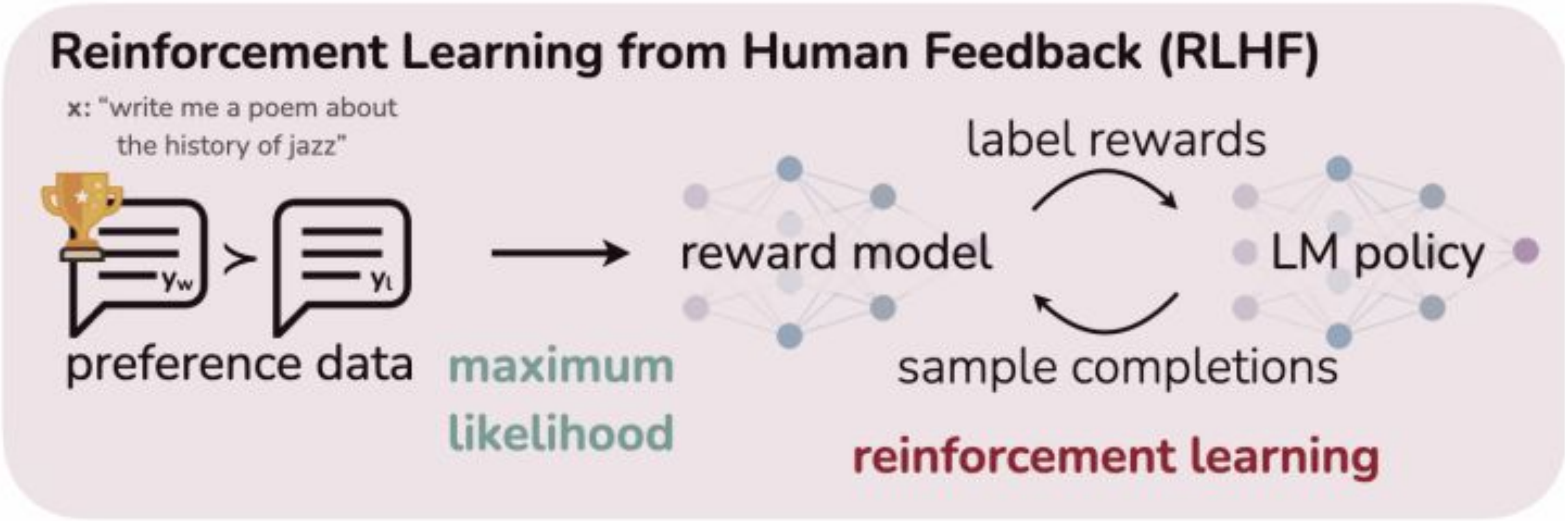


$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\big[\log\sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\big]$$

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\big[r_\phi(x, y)\big] - \beta\mathbb{D}_{\text{KL}}\big[\pi_\theta(y \mid x) \mid\mid \pi_{\text{ref}}(y \mid x)\big]$$

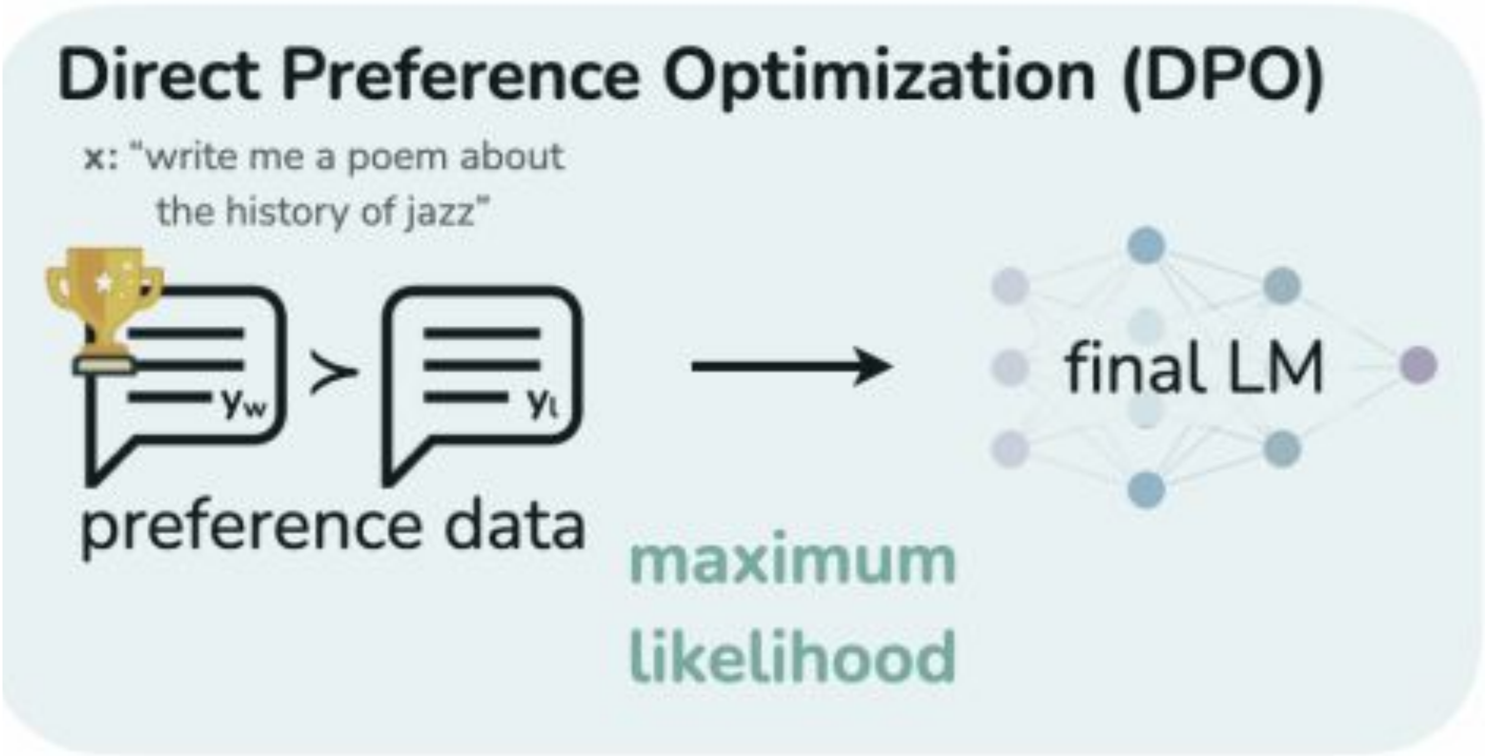
Images from Direct Preference Optimization: Your Language Model is Secretly a Reward Model (Rafailov et al., NeuRIPS 2023)

Online vs Offline Alignment Methods



$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x, y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x)\right]$$



$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)}\right)\right]$$

Images from [Direct Preference Optimization: Your Language Model is Secretly a Reward Model](#) (Rafailov et al., NeuRIPS 2023)

Cross-lingual (X-Lingual) Alignment

- Reward model trained on preference data of language **X** (source)
- Applied to preference tune for language **Y** (target)

Image from Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment (Wu et al., 2024)

Cross-lingual (X-Lingual) Alignment

- Reward model trained on preference data of language **X** (source)
- Applied to preference tune for language **Y** (target)

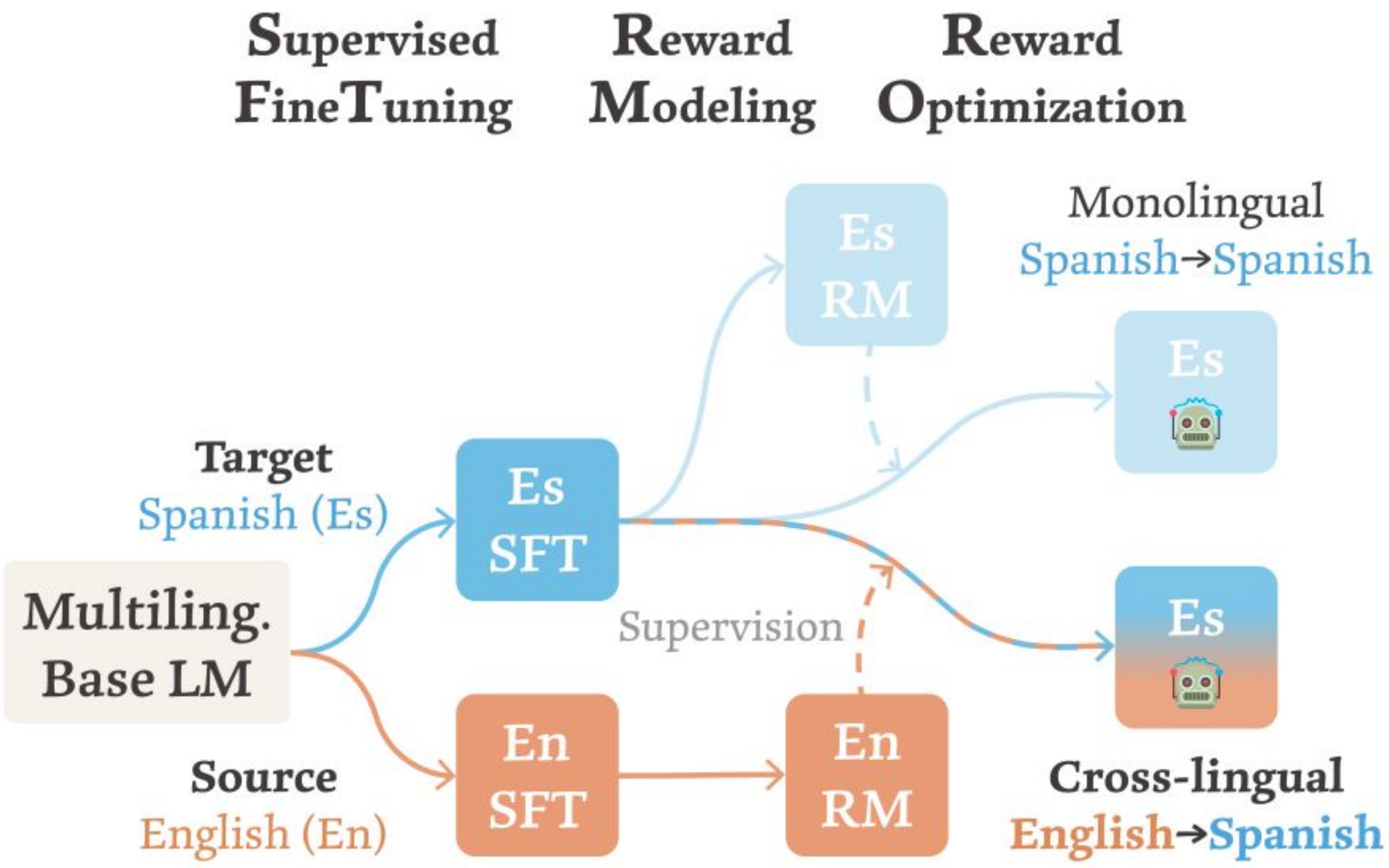


Image from [Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment](#) (Wu et al., 2024)

Cross-lingual Alignment: Does it Work?

(*source*) → (*target*)

- **Evaluation:** Head-to-head win-rates as judged by humans
- **Base SFT model:** mT5-XL
- **Optimization:** Online (PPO)

Cross-lingual alignment sometimes outperforms in-language alignment

Image from [Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment](#) (Wu et al., 2024)

Can’t I Just Translate Source Preference Data?

Cross-lingual

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	52.3	50.8	63.0	66.7	63.0	60.4
En	56.4	55.5	66.1	70.7	67.2	63.1
Es	51.9	51.2	62.4	66.0	64.4	57.5
Ru	48.1	46.5	59.2	63.6	59.0	56.3
Tr	53.3	52.9	62.6	66.6	60.4	59.0
Vi	46.5	48.2	60.0	65.6	62.1	58.0

Table 6: Cross-lingual alignment results using **best-of-n** with $n = 64$, for the **summarization** task, measured in win rate (%) against the target-language SFT model as judged by **PaLM-2-L** (Figure 4).

Translation > Cross-lingual

Translation

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	–	50.0	61.9	66.1	66.1	54.6
En	47.9	–	63.3	64.9	64.5	53.1
Es	50.6	52.9	–	64.1	64.5	59.0
Ru	47.4	51.2	60.3	–	63.3	57.7
Tr	50.6	52.5	61.8	65.6	–	50.8
Vi	42.0	50.8	59.1	64.4	63.6	–

Table 17: Alignment quality using RM trained by translating the source language data into the target language using best-of-n with $n = 64$, for the summarization task, measured in win rate (%) against the target-language SFT model as judged by PaLM-2-L (§5.1).

Can’t say much!!

- English benefits from translation
- Russian (different script) doesn’t transfer well

Cross-lingual Alignment with N languages?

- Cross-lingual works with a language (well mostly!!)
- What if we transfer from more source languages?
- Testbed with various preference mixtures[1]:
    - **En-1:** English-only preference data (50k samples)
    - **ML-5:** 5 language set (en, vi, de, tr & pt) (50k samples, 10k per language)
    - **ML-23:** 23 language set (50k samples, ~2.2k per language)
    - **ML-23*:** 23 language set (230k samples, 10k per language

[1] RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs (Dang et al., 2024)
[2] Command R+ (supports the 23 languages considered for the experiments)

Cross-lingual Alignment with N languages?

- Cross-lingual works with a language (well mostly!!)
- What if we transfer from more source languages?
- Testbed with various preference mixtures[1]:
  - **En-1:** English-only preference data (50k samples)
  - **ML-5:** 5 language set (en, vi, de, tr & pt) (50k samples, 10k per language)
  - **ML-23:** 23 language set (50k samples, ~2.2k per language)
  - **ML-23*:** 23 language set (230k samples, 10k per language)
- For "*ML*" data:
  - Prompts translated from ShareGPT into 22 languages via NLLB
  - Positive Response: Generated multilingual responses to translated prompts via Command R+[2]
  - Negative Response: Generate English response to English prompt via Command and translate
- Tested with offline and online alignment strategies

[1] RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs (Dang et al., 2024)
[2] Command R+ (supports the 23 languages considered for the experiments)

Does Language Diversity help X-lingual Alignment?

- Simulated win-rates with a GPT-4-Turbo

		English Win%	English Loss%	English ΔW-L%			Average 23 Languages Win%	Average 23 Languages Loss%	Average 23 Languages ΔW-L%
DPO	EN-1	52.0	33.5	18.5	DPO	EN-1	43.3	40.6	2.7
	ML-5	50.5	28.5	22.0		ML-5	43.8	39.1	4.7
	ML-23	44.5	36.5	8.0		ML-23	47.0	37.1	9.9
	ML-23*	57.5	31.0	26.5		ML-23*	50.2	39.0	11.2
RLOO	EN-1	47.5	38.5	9.0	RLOO	EN-1	46.4	38.9	7.5
	ML-5	55.5	30.5	25.0		ML-5	54.4	35.8	18.6
	ML-23	53.0	37.0	16.0		ML-23	54.0	38.0	16.0
	ML-23*	53.0	35.0	18.0		ML-23*	53.4	37.0	16.4

Table 3: Open-ended generation (Dolly) win-rates for DPO/RLOO preference optimized Aya models against the original Aya 23 8B on **English (left)** and **averaged over 23 languages (right)**. We report average win-rates on 23 languages for multiple training data mixtures: EN-1 (English Only), ML-5 (5 Languages), and ML-23 (23 Languages). All the data mixtures consist of 50K total training examples with the exception of ML-23*, which includes 230K total training examples. We report results for the best checkpoint across 2 epochs.

Not always for English

Almost always on average across multiple languages

Table from RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs (Dang et al., 2024)

Does More Preference Data Help?

- Simulated win-rates with a GPT-4-Turbo

Yes, it does!

		English Win%	English Loss%	English ΔW-L%	Average 23 Languages Win%	Average 23 Languages Loss%	Average 23 Languages ΔW-L%
DPO	EN-1	52.0	33.5	18.5	43.3	40.6	2.7
	ML-5	50.5	28.5	22.0	43.8	39.1	4.7
	ML-23	44.5	36.5	8.0	47.0	37.1	9.9
	ML-23*	57.5	31.0	26.5	50.2	39.0	11.2
RLOO	EN-1	47.5	38.5	9.0	46.4	38.9	7.5
	ML-5	55.5	30.5	25.0	54.4	35.8	18.6
	ML-23	53.0	37.0	16.0	54.0	38.0	16.0
	ML-23*	53.0	35.0	18.0	53.4	37.0	16.4

Table 3: Open-ended generation (Dolly) win-rates for DPO/RLOO preference optimized Aya models against the original Aya 23 8B on **English (left)** and **averaged over 23 languages (right)**. We report average win-rates on 23 languages for multiple training data mixtures: EN-1 (English Only), ML-5 (5 Languages), and ML-23 (23 Languages). All the data mixtures consist of 50K total training examples with the exception of ML-23*, which includes 230K total training examples. We report results for the best checkpoint across 2 epochs.

Table from RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs (Dang et al., 2024)

What about Languages not in Preference Data?

		Avg. Win %	Unseen Loss %	Langs. ΔW-L%
EN-1	DPO	42.9	40.9	2.0
	RLOO	46.3	39.3	7.3
ML-5	DPO	43.3	39.5	3.8
	RLOO	54.9	35.5	19.4

Table 4: Win-rates for the 22 and 18 languages that are not included in the training data ("unseen") for EN-1 and ML-5 respectively. We observe cross-lingual transfer from preference optimization, with an increased degree of transfer enhanced by multilingual training.

Table from RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs (Dang et al., 2024)

Offline vs Online Alignment

		Average 23 Languages Win%	Loss%	ΔW-L%
DPO	EN-1	43.3	40.6	2.7
	ML-5	43.8	39.1	4.7
	ML-23	47.0	37.1	9.9
	ML-23*	50.2	39.0	11.2
RLOO	EN-1	46.4	38.9	7.5
	ML-5	54.4	35.8	18.6
	ML-23	54.0	38.0	16.0
	ML-23*	53.4	37.0	16.4

		Avg. Unseen Langs. Win %	Loss %	ΔW-L%
EN-1	DPO	42.9	40.9	2.0
	RLOO	46.3	39.3	7.3
ML-5	DPO	43.3	39.5	3.8
	RLOO	54.9	35.5	19.4

Online method is better!

Tables from RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs (Dang et al., 2024)

Challenges

Challenges (The Ones that Made the Cut)

Curse of multilinguality[1,2]

Packing more languages into a model decreases per language performance

Cost of Technology[3]

- GPT* models are behind paid APIs; cost $\propto$ input & generation tokens
- Poor tokenization in non-English languages → more tokens
- More tokens → more latency & money
- Efforts made but far from parity[4,5]

Dialectal Biases[6]

- Whose dialect matters the most?[7,8]
- Whose English?[9,10]

and many more ……

[1] Unsupervised Cross-lingual Representation Learning at Scale (Conneau et al., ACL 2020)
[2] When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages (Chang et al., 2023))
[3] Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models (Ahia et al., EMNLP 2023)
[4] https://cohere.com/blog/command-r-plus-microsoft-azure
[5] https://openai.com/index/hello-gpt-4o/
[6] A Survey of Corpora for Germanic Low-Resource Languages and Dialects (Blaschke et al., NoDaLiDa 2023)
[7] Decolonizing NLP for “Low-resource Languages” (Ògúnrẹ̀mí et al., AI Frameworks Discussion of Abeba Birhane's "Algorithmic Injustice" and Social Impact Articles 2023)
[8] Which Humans? (Atari et al., 2023)
[9] What to do about non-standard (or non-canonical) language in NLP (Plank, KONVENS 2016)
[10] AI makes racist decisions based on dialect (Science, 24 August 2024)

Other Directions

Other Interesting Directions

Multilingual Architectures

- Efficient solutions for the curse of multilinguality
- Adding some language-specific parameters
- E.g.: Adapters[1], Cross-lingual expert models[2]

Tokenization and Vocabulary

- Efficient tokenization methods to reduce costs and latency
- E.g.: Vocab budgeting[6], allocation[7]

Adapting to a New Language

- Increasing support of an **N** language multilingual model to **N+K** languages
- E.g.: Continued pretraining[3], Adapters[4], Efficient Initializations[5]

Data Creation and Verification

- Methods for synthetic data generation[8] and verification of labeled data[9]

[1] MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer (Pfeiffer et al., EMNLP 2020)
[2] Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models (Blevins et al., 2024)
[3] How to Adapt Your Pretrained Multilingual Model to 1600 Languages (Ebrahimi & Kann, ACL-IJCNLP 2021)
[4] BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting (Yong et al., ACL 2023)
[5] OFA: A Framework of Initializing Unseen Subword Embeddings for Efficient Large-scale Multilingual Continued Pretraining (Liu et al., Findings 2024)
[6] XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models (Liang et al., EMNLP 2023)
[7] Tokenization Impacts Multilingual Language Modeling: Assessing Vocabulary Allocation and Overlap Across Languages (Limisiewicz et al., Findings 2023)
[8] Multilingual Arbitrage: Optimizing Data Pools to Accelerate Multilingual Progress (Odumakinde et al., 2024)
[9] Verifying Annotation Agreement without Multiple Experts: A Case Study with Gujarati SNACS (Mehta & Srikumar, Findings 2023)

References & Future Readings

Inequalities in Technology across Languages

- Breaking the unwritten language barrier: The bulb project (Adda et al., 2016)
- The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi et al., ACL 2020)
- Global predictors of language endangerment and the future of linguistic diversity(Bromham et al., 2021, Nature Ecology&Evolution)
- Systematic Inequalities in Language Technology Performance across the World's Languages (Blasi et al., ACL 2022)
- Which Humans? (Atari et al., 2023)
- Decolonizing NLP for "Low-resource Languages" (Ògúnrẹ̀mí et al., AI Frameworks Discussion of Abeba Birhane's "Algorithmic Injustice" and Social Impact Articles 2023)
- Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models (Ahia et al., EMNLP 2023)
- Abundance of words versus poverty of mind: the hidden human costs co-created with LLMs (Vuong and Ho, AI & Society 2024)

Multilingual Language Models

- mBART: Multilingual Denoising Pre-training for Neural Machine Translation (Liu et al., TACL 2020)
- mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)
- BLOOM: A 176B-Parameter Open-Access Multilingual Language Model (BigScience, 2022)
- xGLM: Few-shot Learning with Multilingual Generative Language Models (Lin et al., 2023)
- Glot500-m: Glot500: Scaling multilingual corpora and language models to 500 languages (Imani et al., 2023)
- PolyLM: An Open Source Polyglot Large Language Model (Wei et al., 2023)
- BLOOMZ: Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., ACL 2023)
- mT0: Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., ACL 2023)
- Okapi series: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback (Lai et al., 2023)
- mGPT: Few-Shot Learners Go Multilingual (Shliazhko et al., TACL 2024)
- Aya-101: Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., 2024)
- MALA-500: Massive Language Adaptation of Large Language Models (Lin et al., 2024)
- Aya-23: Open Weight Releases to Further Multilingual Progress (Aryabumi et al., 2024)

References & Future Readings

Multilingual Pre-training

- mC4: mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., 2021)
- ROOTS: The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset (Laurençon, NeurIPS 2022)
- Glot2000-c & Glot500-c: Scaling multilingual corpora and language models to 500 languages (Imani et al., 2023)

Multilingual Instruction-Tuning

- Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks (Wang et al., EMNLP 2022)
- Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback (Lai et al., 2023)
- xP3 & xP3mt: Crosslingual generalization through multitask finetuning (Muennighoff et al., ACL 2023)
- xP3x: Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model (Üstün et al., 2024)
- Aya Dataset & Collection: Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning (Singh et al., ACL 2024)
- Multilingual Instruction Tuning With Just a Pinch of Multilinguality (Shaham et al., Findings 2024)

Multilingual Preference and Safety Alignment

- Multilingual Jailbreak Challenges in Large Language Models (Deng et al., 2023)
- The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts (Shen et al., 2024)
- Having Beer after Prayer? Measuring Cultural Bias in Large Language Models (Naous et al., 2024)
- All Languages Matter: On the Multilingual Safety of LLMs (Wang et al., Findings 2024)
- From One to Many: Expanding the Scope of Toxicity Mitigation in Language Models (Ermis et al., Findings 2024)
- Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment (Wu et al., 2024)
- Preference Tuning For Toxicity Mitigation Generalizes Across Languages (Li et al., 2024)
- RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs (Dang et al., 2024)
- PolygloToxicityPrompts: Multilingual Evaluation of Neural Toxic Degeneration in Large Language Models (Jain et al., 2024)
- The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm (Aakanksha et al., 2024)

References & Future Readings

Curse of Multilinguality and Architectural Solutions

- Unsupervised Cross-lingual Representation Learning at Scale (Conneau et al., ACL 2020)
- MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer (Pfeiffer et al., EMNLP 2020)
- MAD-G: Multilingual Adapter Generation for Efficient Cross-Lingual Transfer (Ansell et al., Findings 2021)
- Efficient Test Time Adapter Ensembling for Low-resource Language Varieties (Wang et al., Findings 2021)
- Cross-lingual Few-Shot Learning on Unseen Languages (Winata et al., AACL-IJCNLP 2022)
- Lifting the Curse of Multilinguality by Pre-training Modular Transformers (Pfeiffer et al., NAACL 2022)
- BAD-X: Bilingual Adapters Improve Zero-Shot Cross-Lingual Transfer (Parović et al., NAACL 2022)
- Hyper-X: A Unified Hypernetwork for Multi-Task Multilingual Transfer (Üstün et al., EMNLP 2022)
- When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages (Chang et al., 2024)
- Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models (Blevins et al., arXiv 2024)

NLP for Indigenous Languages

- Local Languages, Third Spaces, and other High-Resource Scenarios (Bird, ACL 2022)
- Not always about you: Prioritizing community needs when developing endangered language technology (Liu et al., ACL 2022)
- "It's how you do things that matters": Attending to Process to Better Serve Indigenous Communities with Language Technologies (Cooper et al., EACL 2024)
- Modeling the Sacred: Considerations when Using Religious Texts in Natural Language Processing (Hutchinson, Findings 2024)
- Must NLP be Extractive? (Bird, 2024)
- Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences (Pinhanez et al., 2024)

References & Future Readings

Adapting to New Languages

- How to Adapt Your Pretrained Multilingual Model to 1600 Languages (Ebrahimi & Kann, ACL-IJCNLP 2021)
- Phylogeny-Inspired Adaptation of Multilingual Models to New Languages (Faisal & Anastasopoulos, AACL-IJCNLP 2022)
- Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation (Wang et al., ACL 2022)
- Don't Stop Fine-Tuning: On Training Regimes for Few-Shot Cross-Lingual Transfer with Multilingual Language Models (Schmidt et al., EMNLP 2022)
- Cross-lingual Continual Learning (M'hamdi et al., ACL 2023)
- Mini-Model Adaptation: Efficiently Extending Pretrained Models to New Languages via Aligned Shallow Training (Marchisio et al., Findings 2023)
- BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting (Yong et al., ACL 2023)
- OFA: A Framework of Initializing Unseen Subword Embeddings for Efficient Large-scale Multilingual Continued Pretraining (Liu et al., Findings 2024)

Miscellaneous

- How Vocabulary Sharing Facilitates Multilingualism in LLaMA? (Yuan et al., 2023)
- Tokenization Impacts Multilingual Language Modeling: Assessing Vocabulary Allocation and Overlap Across Languages (Limisiewicz et al., Findings 2023)
- XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models (Liang et al., EMNLP 2023)
- Do Multilingual Language Models Think Better in English? (Etxaniz et al., NAACL 2024)
- Do Llamas Work in English? On the Latent Language of Multilingual Transformers (Wendler et al., ACL 2024)
- How Does Quantization Affect Multilingual LLMs? (Marchisio et al., 2024)
- Multilingual Arbitrage: Optimizing Data Pools to Accelerate Multilingual Progress (Odumakinde et al., 2024)

The End.