

Data and Model Centric Approaches for Expansion of Large Language Models to New languages

Anoop Kunchukuttan, Raj Dabre, Rudra Murthy V,
Mohammed Safi Ur Rahman Khan, Thanmay Jayakumar

<https://ai4bharat.github.io/multilingual-llm-expand-emnlp-2025>

8th November 2025, EMNLP, Suzhou, China



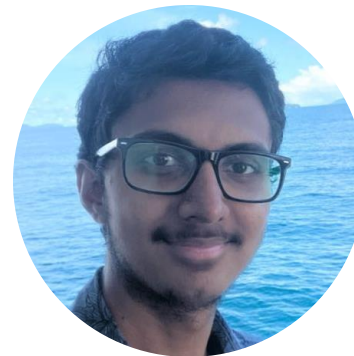
AI4Bharat



Tutorial Presenters



Mohammed Safi Ur Rahman Khan
WSAI, IIT Madras



Thanmay Jayakumar
WSAI, IIT Madras



Anoop Kunchukuttan
Microsoft Speech India



Raj Dabre
IIT Madras



Rudramurthy V
IBM Research India

Tutorial Outline

- Introduction
 - The Multilingual LLM Challenge
- Extending LLMs to new languages
 - Vocabulary Expansion
 - Continued Pre-training
 - **<BREAK>**
 - Post-Training
- Multilingual Evaluation
- Summary & Future Directions

[BIBLIOGRAPHY](#) (identify paper by *[PAPER_KEY]* on the slides)

Tutorial Outline

- **Introduction**

- The Multilingual LLM Challenge
- Extending LLMs to new languages
 - Vocabulary Expansion
 - Continued Pre-training
 - Post-Training
- Multilingual Evaluation
- Summary & Future Directions

What are Large Language Models?

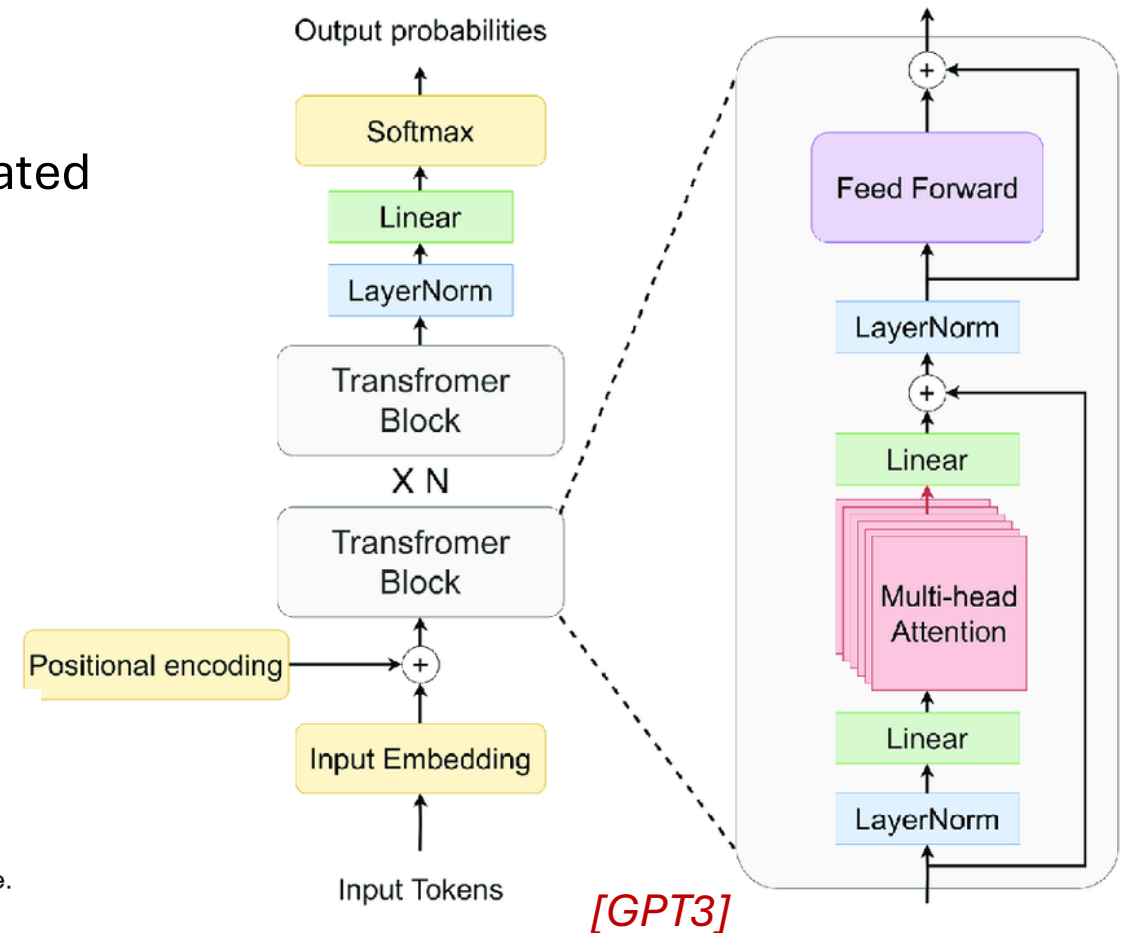
- Typically, transformer decoder models
- They generate text by looking at only previously generated text (*auto-regressive*)
- Trained on a **self-supervised** task
 - Next word prediction task
 - Large amount of text data
 - Large Models
- **In-context learning capability**

Instruction Tell me the sentiment of this review

Example The movie begins The plot is engaging, thoroughly enjoyable.
The movie is **great**

Oh, how can such a fine cast produce such a terrible performance..... A total waste of time.
The movie is **pathetic**

Prompt It is just a rehash of old movies
The movie is <MASK>



[InstructGPT]

- Finetuning on (relatively) small supervised and preference data to align instructions and values

Current LLMs vs. older generation (BERT/BART/XLM-R)

Current

- Decoder-only (NLU and NLG)
- Causal LM training objectives
- Every task is just text completion
- In-context learning & Post-training
- Large model size (GPT3: 175B params)
- Trained on large corpora (Qwen3: 36T tokens)

Old Generation

- Encoder-only (NLU), Enc-dec (NLG)
- Denoising objectives (MLM, DAE)
- Classification/text generation
- Per-task finetuning
- Small models (largest mT5: ~13B)
- Modest amount of data (~BERT: 137B tokens)

What makes LLMs exciting?

- Inherent knowledge
- Multi-task abilities and task-composition
- Little task-specific training required
- Help in writing , creativity, brainstorming, data generation, explanation, etc.
- Reasoning Capabilities

	Q&A Answer questions based on existing knowle...		Grammar correction Corrects sentences into standard English.
	Summarize for a 2nd grader Translates difficult text into simpler concep...		Natural language to OpenAI API Create code to call to the OpenAI API usin...
	Text to command Translate text into programmatic commands.		English to other languages Translates English text into French, Spanish...
	Natural language to Stripe API Create code to call the Stripe API using nat...		SQL translate Translate natural language to SQL queries.
	Parse unstructured data Create tables from long form text		Classification Classify items into categories via example.
	Python to natural language Explain a piece of Python code in human un...		Movie to Emoji Convert movie titles into emoji.
	Calculate Time Complexity Find the time complexity of a function.		Translate programming languages Translate from one programming language ...
	Advanced tweet classifier Advanced sentiment detection for a piece o...		Explain code Explain a complicated piece of code.

Maths

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

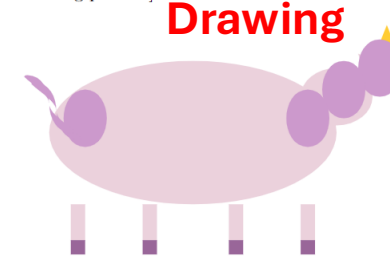
Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s

Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 4]$, $y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ to be "allie". I want the x -axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in "continuous time". Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

Prompt: Draw a unicorn in TikZ.

GPT-4: [Produces L^AT_EX compiling to following picture.]

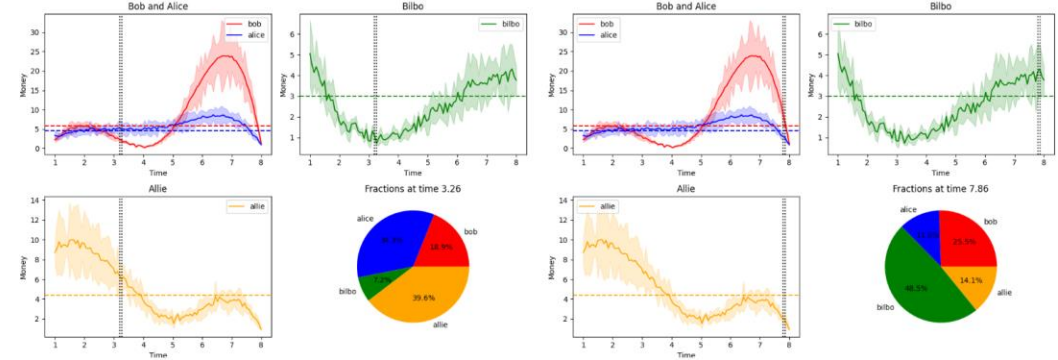


Drawing

Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 4]$, $y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ to be "allie". I want the x -axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in "continuous time". Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

GPT-4: [We give two snapshots from the animation resulting from the produced code]



Plotting

GPT-4: Explanations

This is a conversation between a user and an artificial intelligence agent that follows instructions or answers questions. The AI always follows the instructions correctly. --- User: Can you compose a short tune (say four to eight bars) using ABC notation? AI: There are many possible ways to compose a short tune in abc notation, but here is one example: X:1 T:Example Tune M:4/4 L:1/8 K:C |:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4- G2z2 :| |:c2de f2ed | c2de f2f2 | g2fe d2cB | c4- c2z2 :| This tune has two four-bar phrases, each repeated once, forming an A part and a B part. The key is C major.

GPT-4:

User: After the first bar G2AB c2BA, you chose G to start the next bar. Give an example of a change to the notes in the first bar that would have caused you to pick D instead of G as the first note in the second bar. AI:One possible change to the notes in the first bar that would have caused me to pick D instead of G as the first note in the second bar is:

|:G2AB c2Bc | ...

This change creates a stronger cadence on the C note at the end of the first bar, which is the fifth degree of the scale. This suggests a modulation to the dominant key of G major, which has D as its fifth degree. Therefore, starting the second bar with D would create a smooth transition to the new key and imply a harmonic progression of V-I.

Prompt:

```
struct mystruct_A { char a; int b; char c; } x;
struct mystruct_B { int b; char a; char c; } y;

#include <stdio.h>
int main(void) {
    printf("Size of x: %zu bytes\n", sizeof(x));
    printf("Size of y: %zu bytes\n", sizeof(y));
    return 0;
}
```

Possible Output:

GPT-4:

There are different ways to implement and align structures in memory, so the exact output may vary depending on the compiler, platform, and options used. However, one possible output is:

Size of x: 12 bytes
Size of y: 8 bytes

This output assumes that the compiler uses 4-byte alignment for structures, meaning that each structure member starts at an offset that is a multiple of 4 bytes from the beginning of the structure. This may result in some padding bytes being inserted between or after the members to maintain alignment. For example, in x, there may be 3 padding bytes after a and 3 padding bytes after c, while in y, there may be no padding at all. The size of a structure is the sum of the sizes of its members plus any padding bytes.

Understanding Programs

Reasoning

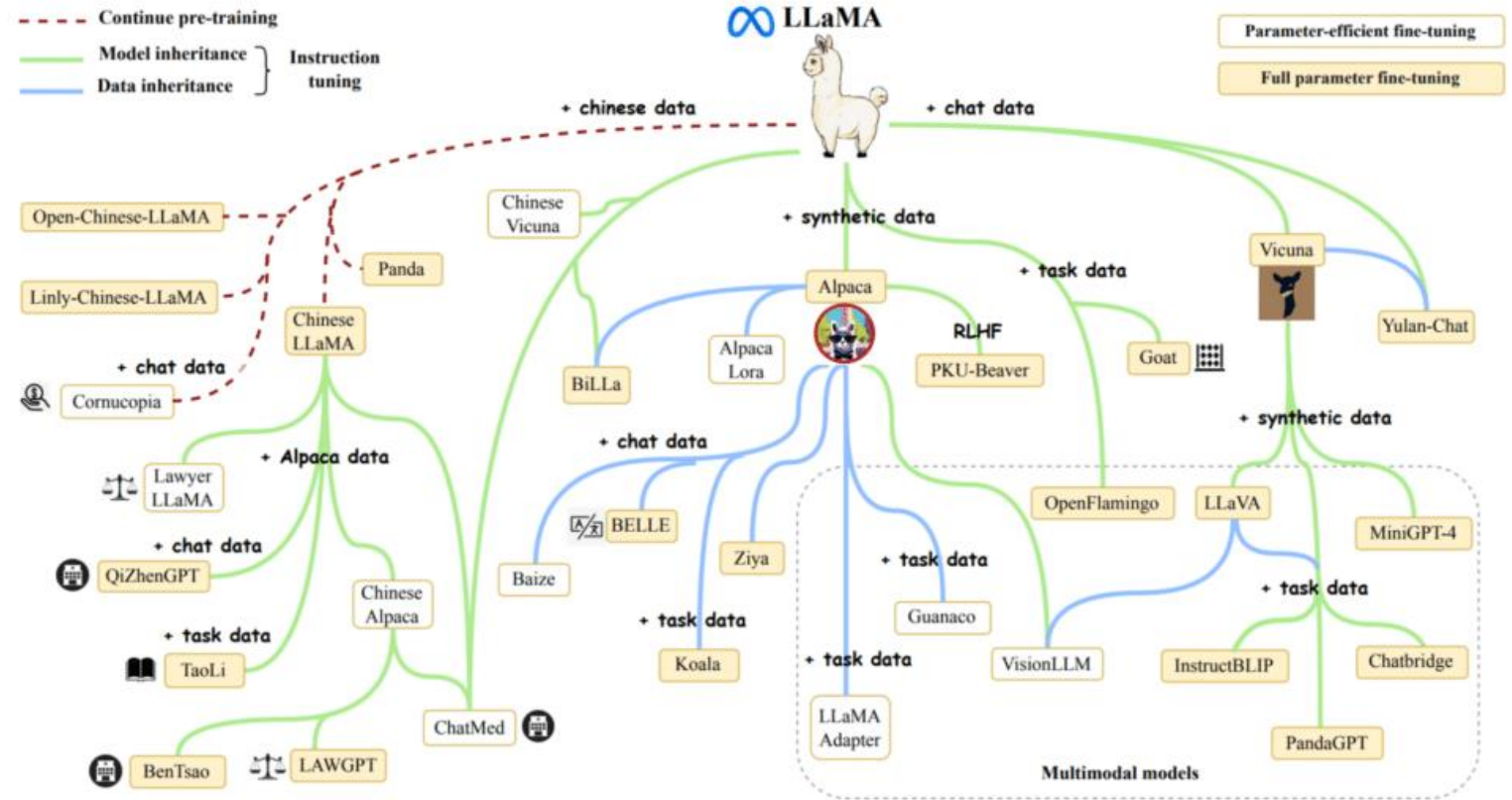
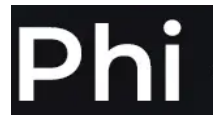
Why is 31 a prime number?

Reasoned about number 31 being prime for 4 seconds ✓

Alright, let's go through this. A prime number is an integer greater than 1, divisible only by 1 and itself. So why is 31 prime? Its only divisors are 1 and 31. I check for divisibility — it's not divisible by 2 ($31/2 = 15.5$, remainder 1), not by 3 ($31/3 = 10.33$, remainder 1), and not divisible by 5 ($31/5 = 6.2$, remainder 1). Since no other primes divide it, 31 is prime!

Since 31 has no divisors other than 1 and itself, it is prime. To determine this, we check possible divisors up to $\sqrt{31}$, approximately 5.6. The divisors to check are 2, 3, and 5, and none divide evenly into 31 — $31/2 = 15.5$, $31/3 \approx 10.33$, $31/5 = 6.2$. Since 31 isn't divisible by these, it confirms that 31 is prime. The fundamental theorem of arithmetic also supports that 31 cannot be factored further.

Explosion of LLMs ... but mostly limited to English



Benefits of LLMs are mostly limited to English

Results on XNLI

Language	Cat.	ChatGPT	
		(en)	(spc)
English	H	70.2	70.2
Russian	H	60.8	45.4
German	H	64.5	51.1
Chinese	H	58.2	35.5
French	H	64.8	42.2
Spanish	H	65.8	47.4
Vietnamese	H	55.4	44.8
Turkish	M	57.1	37.1
Arabic	M	55.3	22.3
Greek	M	55.9	54.5
Thai	M	44.7	11.5
Bulgarian	M	59.7	44.6
Hindi	M	48.8	5.6
Urdu	L	43.7	6.3
Swahili	X	50.3	40.8

Results on X-CSQA

Language	Code	Cat.	ChatGPT	
			(en)	(tgt)
English	en	H	75.0	75.0
Russian	ru	H	50.2	53.5
German	de	H	52.6	61.0
Chinese	zh	H	50.2	42.5
Japanese	jp	H	41.9	43.0
French	fr	H	50.5	61.7
Spanish	es	H	53.3	62.5
Italy	it	H	50.6	55.9
Dutch	nl	H	52.9	60.4
Polish	pl	H	35.2	51.1
Portuguese	pt	H	49.5	59.2
Vietnamese	vi	H	42.3	47.9
Arabic	ar	M	49.4	47.3
Hindi	hi	M	41.1	38.6
Urdu	ur	L	34.7	24.5
Swahili	sw	X	35.6	46.6
Average			47.8	51.9

Language	Cat.	ChatGPT(en)	
		EM	F1
English	H	56.0	74.9
Russian	H	30.2	49.1
German	H	45.9	65.8
Chinese	H	37.1	42.3
Spanish	H	41.8	65.8
Vietnamese	H	36.1	57.3
Turkish	M	34.5	56.4
Arabic	M	32.0	50.3
Greek	M	29.7	45.0
Thai	M	31.2	43.4
Hindi	M	17.5	37.8
Average		35.6	53.5

Results on XQuad QnA

Model	EN	AVG
Qwen2.5-32B-Instruct	38.43	29.41
Gemma3-27B-IT	50.55	44.88
QwQ-32B	79.43	74.69
Deepseek-R1	78.81	75.72
o3-mini	82.18	79.90

Performance on MMATH

Lang.	ChatGPT		NLLB	
	BLEU	chrF	BLEU	chrF
srp_Cyrl	1.36	3.26	43.4	59.7
kon_Latn	0.94	8.50	18.9	45.3
tso_Latn	2.92	15.0	26.7	50.0
kac_Latn	0.04	2.95	14.3	37.5
nso_Latn	3.69	16.7	26.5	50.8
jpn_Jpan	28.4	32.9	20.1	27.9
nno_Latn	37.1	58.7	33.4	53.6
zho_Hans	36.3	31.0	26.6	22.8
zho_Hant	26.0	24.4	12.4	14.0
acm_Arab	28.2	44.7	11.8	31.9

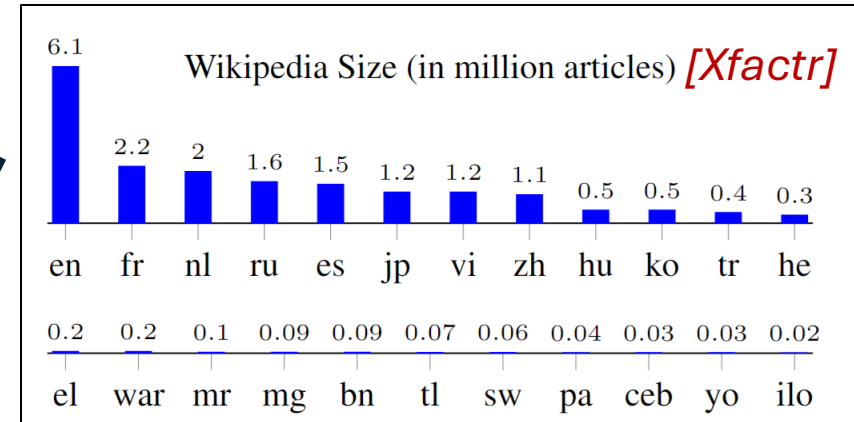
Performance on translation High vs low resource

- Significant gap between English and other languages on multiple tasks
- High-resource and Latin script languages can give good performance on GPT
- Poor performance on low-resource languages
- Translate-test is a strong baseline
- Open-source models lag behind GPT models → they are very English heavy

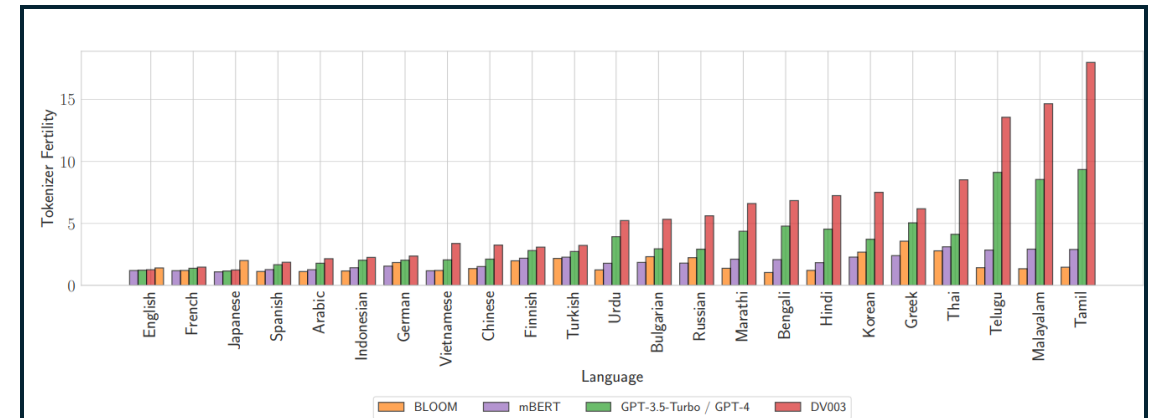
[BUFFET, MEGA, ChatGptMT, ChatGptMLing, MMATH]

Why do LLMs lag behind for other languages?

- Lack of
 - Pre-training data
 - Token representation
 - Instruction tuning data
 - Human preference data
 - Reasoning data
- Inability to transfer from English
- Limitations of Translate-Test



*Most LLMs
trained on <10%
non-English data*



Fertility → number of tokens per word

High fertility → low-efficiency, suboptimal representations

[BUFFET, MEGA, ChatGptMT]

Why extend English LLMs to new languages?

- Very few high-quality multilingual open LLMs
- Large-scale compute needed for training multilingual models from scratch
- Extending existing LLMs can be compute-efficient
- Leverage strong English capabilities of base LLMs
- English LLMs are at the cutting edge with regular updates

Major projects extending English LLMs:

ChineseLLama, OpenHathi, Sarvam-M, RomanSetu, SeaLLM, Sea-LION, Sailor, ALMA, X-ALMA, Tower series

Do English-dominant LLMs have some inherent multilingual capabilities?

Yes, to some extent ...

Why? – during training they might have been exposed to some non-English data

- Documents with multiple languages
- Incorrect LID
- Increasingly some representation of non-English data e.g. Gemma3, LLama4, Qwen3

How good are the multilingual capabilities?

- Might be ok at language understanding e.g. *classification, sentiment analysis*
- Bad at generation
- Better on Latin script languages
- Languages with better pre-training representation perform better

How do English LLMs achieve multilingual capabilities?

- *Do LLMs think in English?*
- *Do LLM use English as a pivot for decision making?*

Bottom layers: Feature learning

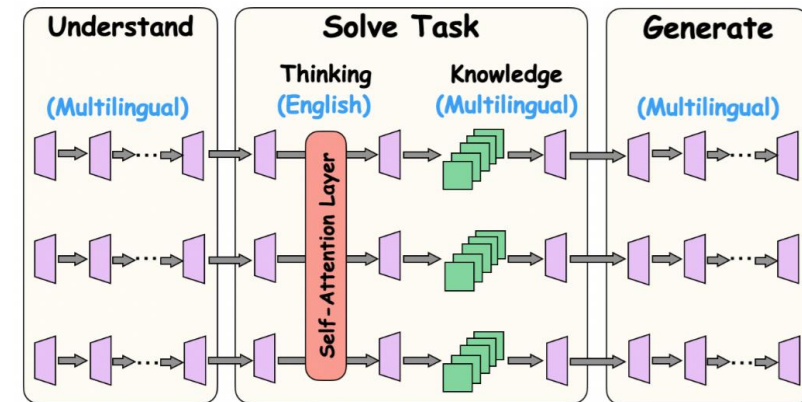
*Middle layers: Concept mapping to language tokens
(with English bias), task solving*

Top layers: Language generation in target language

There are language-specific neurons (mainly concentrated in the top and bottom layers)

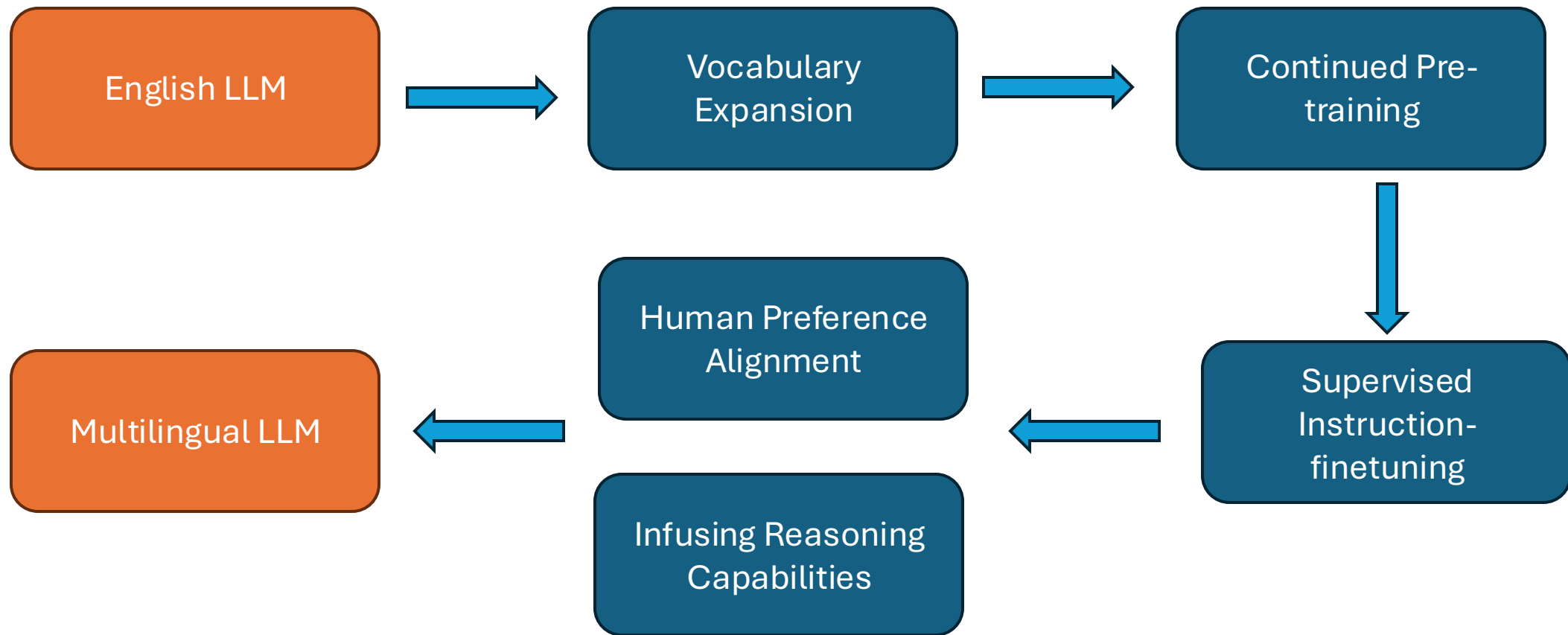
The central question in building multilingual LLM is to bring representations of English and other languages closer to achieve good cross-lingual transfer

[LmaLatent,PNLD,LSP,SharingNeurons]



Output	文	:	"	花
31	文	:	"	花
29	文	:	"	花
27	文	:	_flower	花
25	文	:	_flowe...	_flowe...
23	文	:	"	_flowe...
21	文	:	_flowe...	_flowe...
19	文	:	"	_flowe...
17	eval	:	"	<0xE5>
15	ji	:	"	ψ
13	i	_vac	ols	_bore
11	eda	eda	_Als	abei
9	eda	ná	_Als	_hel
7	iser	arie	◀	arias
5	npa	orr	◀	arias
3	心	ures	_Bedeut	arda
1	_beskre	化	Portail	_Kontr...
	中	文	:	"

Extending English LLMs to Non-English Languages

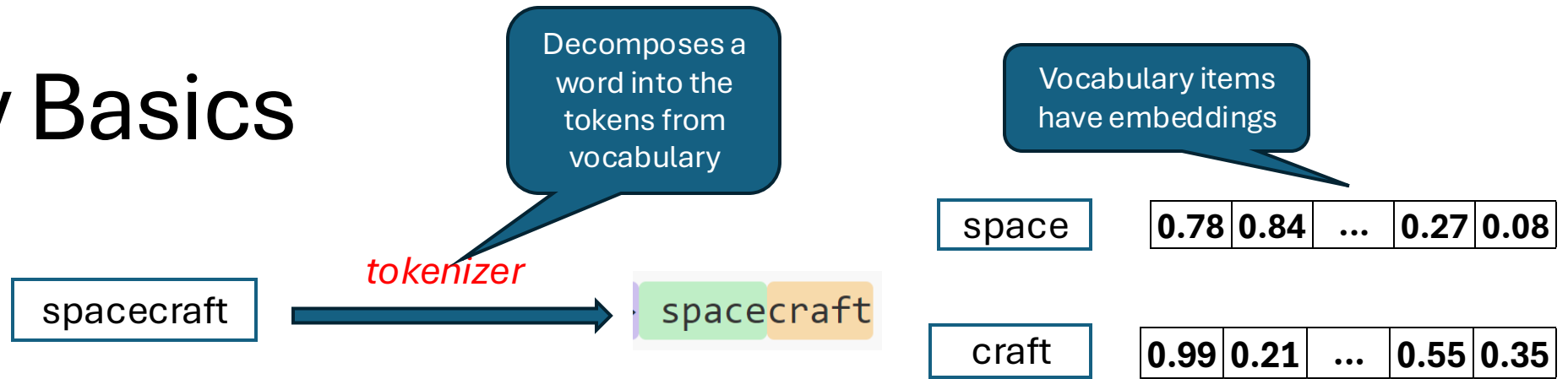


Tutorial Outline

- Introduction
 - The Multilingual LLM Challenge
- Extending LLMs to new languages
 - **Vocabulary Expansion**
 - Continued Pre-training
 - Post-Training
- Multilingual Evaluation
- Summary & Future Directions

Vocabulary Expansion

Vocabulary Basics



<s> Gaganyaan is an Indian crewed orbital spacecraft intended to be the formative spacecraft of the Indian Human Spaceflight Programme.

Vocabulary: Set of **tokens** (basic I/O units)

LLM Vocabulary Properties

- **Finite** vocabulary size
- **Subword** units: basic units are smaller than words
- **Open** vocabulary: all words can be defined as concatenation of subwords

What if vocabulary is under-represented?

<s> गगनयान <0xE0><0xA4><0x8F>क भारतीय चालक दल कक्षीय अंतरिक्ष यान है जिसका <0xE0><0xA4><0x89>द्देश्
य भारतीय मानव अंतरिक्ष <0xE0><0xA4><0x89>डान कार्यक्रम का प्रारंभिक अंतरिक्ष यान होना है।

Fertility = Average number of tokens per word

Unknown characters (BPE-based vocab)	UNK vocab item
Fallback to known characters (BPE-based vocab)	High Fertility
Fallback to bytes (Byte BPE-based vocab)	Even Higher Fertility

High fertility ➔

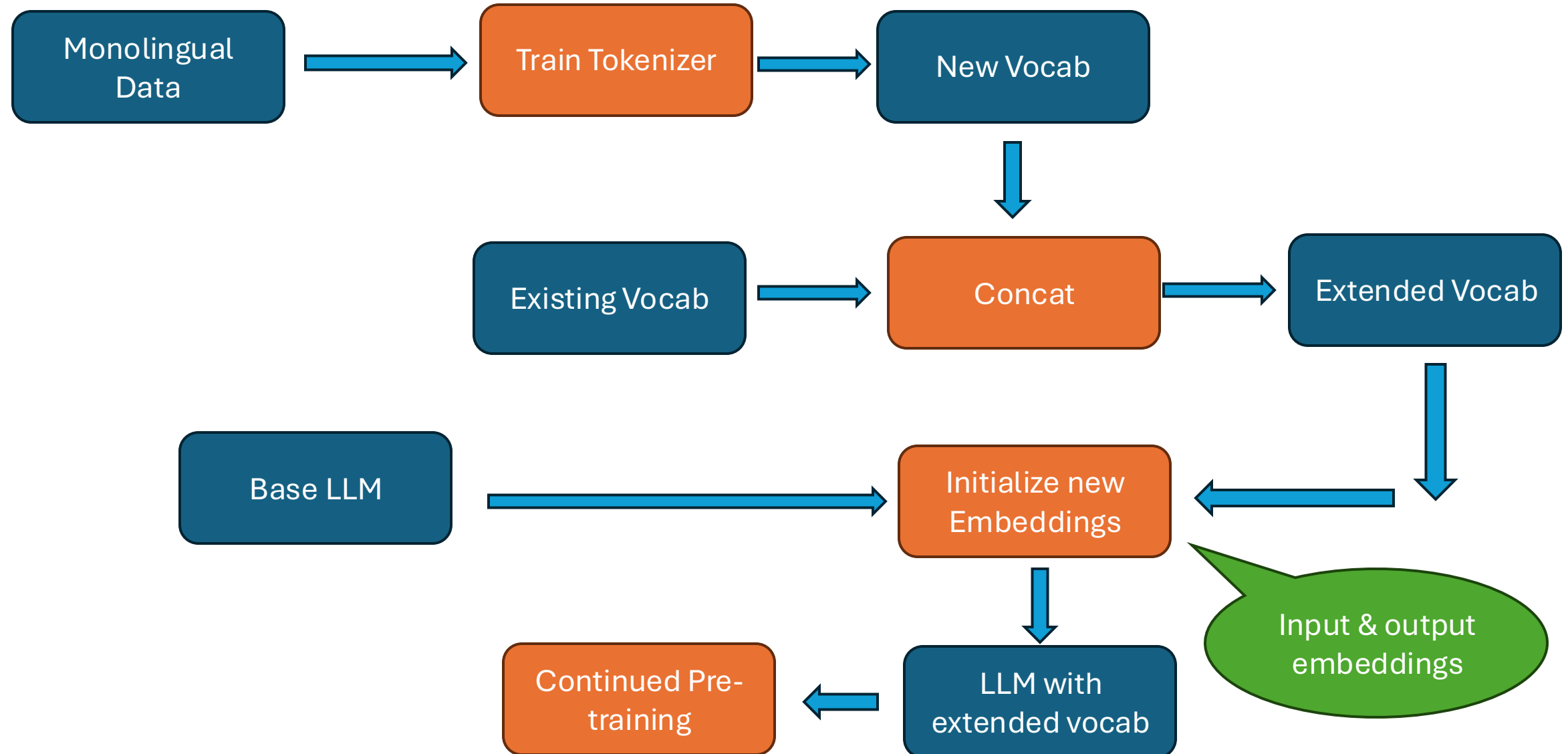
Noisier embeddings and lesser semantic cohesion

More memory consumption

More decoding time

Limit on longest processable sequence

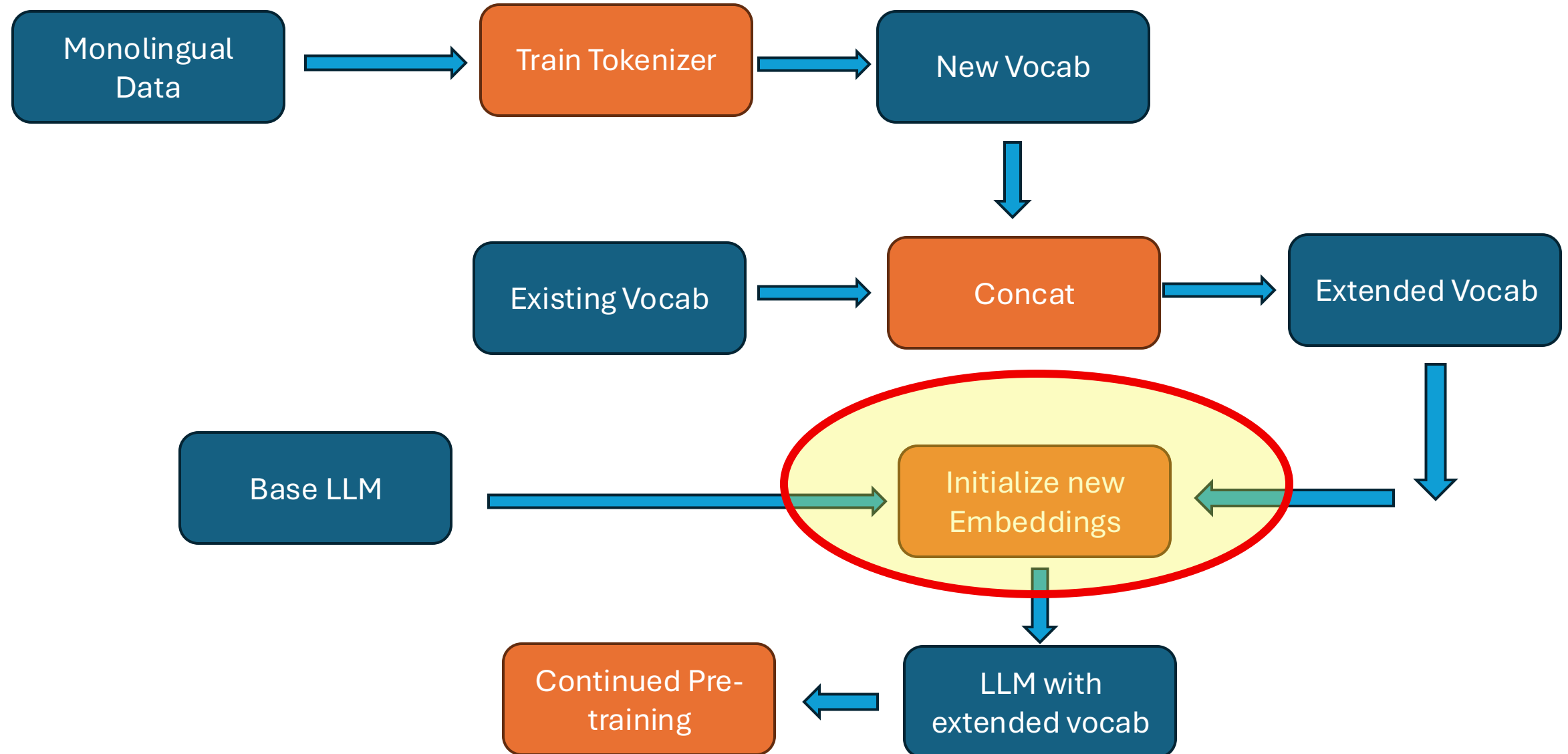
How to extend tokenizer vocabulary?



What do we want from a good vocabulary expansion technique?

- Achieve **improved fertility** → efficient inference
- **Optimize compute** required for continued pre-training
- **Optimize data** required for continued pre-training
- **Task performance goals**
 - Retain performance on original languages (generally English)
 - Improve performance on the language of interest
 - Achieve performance close to joint training of multilingual model
- Enable better **cross-lingual transfer**

How to extend tokenizer vocabulary?



Some simple embedding initialization methods

- Randomly initialize all embeddings
 - Copy embeddings from original vocab to new vocab items via random assignment
 - Copy embeddings of shared vocabulary items
-
- *Changes existing vocab's probability distribution*
 - *Incorrect generation in existing language*
 - *Large convergence time*

[FVT, EfficientCVA, SmallDataCVA, ConstrainedW2V]

Average Initialization

[AveInit, ConstrainedW2V, ExpandChoices, SmallDataCVA]

A simple solution: Initialize new tokens to average of embeddings of existing tokens

- Low KL-divergence between pre- and post-expansion LMs for existing vocabulary
- Greedy decoding with prefix of existing tokens will result in output from existing tokens
- **A general result:** the above applies if new embeddings are in the convex hull of existing embeddings

A practical solution: We want to avoid all new embeddings been initialized to same value

1. Add small random noise to the average embeddings
2. Decompose new tokens using existing tokenizer – average of these embeddings.

Strong baseline [ConstrainedW2V, SmallDataCVA]

Weighted Average Initialization [WECHSEL]

- Target token embeddings (e^t) as weighted average of source token embeddings (e^s)
- Weights based on source-target token similarities from auxiliary static embeddings ($s_{x,y}$)

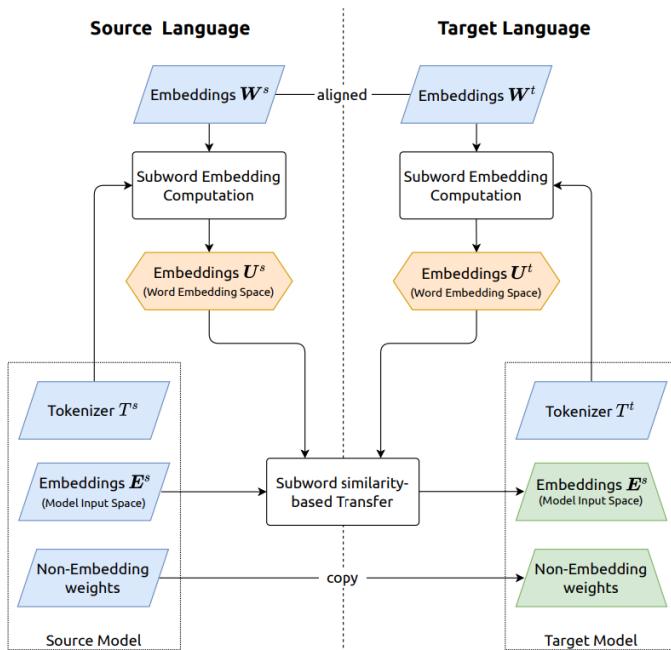


Figure 1: Summary of our **WECHSEL** method. We show **inputs**, **intermediate results** and **outputs**.

$$s_{x,y} = \frac{u_x^t u_y^s T}{\|u_x^t\| \|u_y^s\|}$$

$$e_x^t = \frac{\sum_{y \in \mathcal{J}_x} \exp(s_{x,y}/\tau) \cdot e_y^s}{\sum_{y' \in \mathcal{J}_x} \exp(s_{x,y'}/\tau)}$$

Ways of auxiliary obtaining static embeddings

1. Using subword-aware word embeddings (like FastText) [WECHSEL]

$$u_x = \sum_{g \in \mathbb{G}(x)} w_g$$

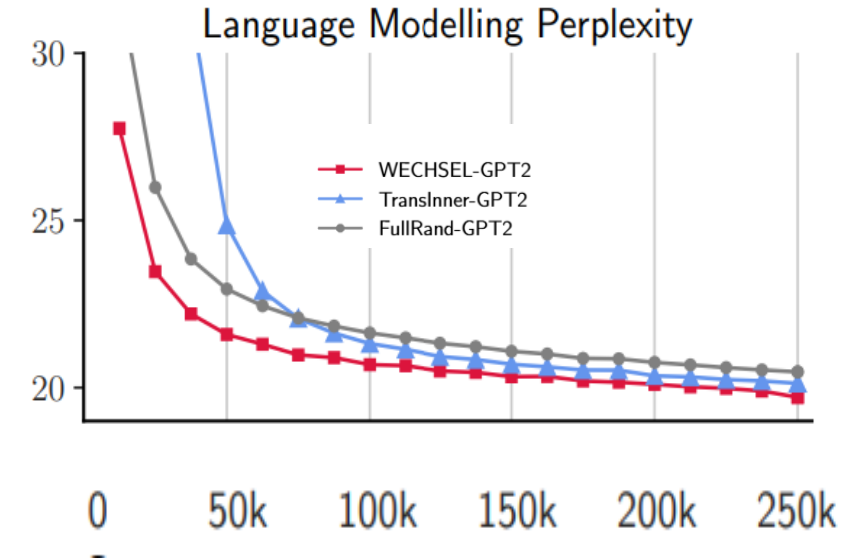
2. From weighted average of word-embeddings containing token [OFA, WECHSEL]

$$u_x = \frac{\sum_{v \in \mathbb{V}(x)} w_v \cdot f_v}{\sum_{v \in \mathbb{V}(x)} f_v}$$

3. Directly learn the static subword embeddings [FOCUS]

Weighted Average Initialization (2) [WECHSEL]

Model	Score@0			Score@25k			Score@250k		
	NLI	NER	Avg	NLI	NER	Avg	NLI	NER	Avg
WECHSEL-RoBERTa	<u>78.25</u>	<u>86.93</u>	<u>82.59</u>	<u>81.63</u>	<u>90.26</u>	<u>85.95</u>	<u>82.43</u>	<u>90.88</u>	<u>86.65</u>
TransInner-RoBERTa	60.86	69.57	65.21	65.49	83.82	74.66	81.75	90.34	86.04
FullRand-RoBERTa	55.71	70.79	63.25	69.02	84.24	76.63	75.28	89.30	82.29
XLM-R _{Base} (Final)	79.25	89.48	84.37						



Better Initialization, but still needs continued pre-training

After Continued Pre-training

Faster convergence vs. random baselines for

- LM perplexity
- Downstream performance

WECHSEL Extensions

OFA (One-for-All): *multilingual vocabulary, need to handle large vocab* [OFA]

- **Factorize and reduce embedding dimension** (inspired from ALBERT)
- Source embedding factorization with SVD
 - Co-ordinates: language-dependent
 - Primitives: language-independent
- Projection of source co-ordinates to target co-ordinates like WECHSEL

FOCUS:

*Target token embeddings as weighted average of **overlapping** source/target token embeddings* [FOCUS]

Learning the weights of the linear combination

Heuristic approaches like WECHSEL, OFA, etc.

- Need cross-lingual embeddings or bilingual dictionaries
- Do not optimize for any specific objective

Can we learn the weight matrix **W** that represents the linear combination weights?

$$\mathbf{E}^t = \mathbf{W} \mathbf{E}^s$$

What should be the learning objective?

Word2Vec objective *[ConstrainedWord2Vec]*

Model learning new token embeddings as a Word2Vec problem (like skipgram)

with the following constraints:

- Embeddings of existing tokens are frozen
- Embeddings of new tokens are strictly expressed as linear combination of existing tokens

In-expensive, no bilingual embeddings required

LM Objective *[VocADT]*

Directly optimize for the LM objective i.e. next token prediction

- Freeze all original LLM parameters and update only the weight matrix **W**
- Initialize the weights **W** to copy shared vocabulary embeddings
- Auxiliary objective to prevent shared vocabulary embeddings from deviating too far

Directly optimizes for end objective, more computationally expensive

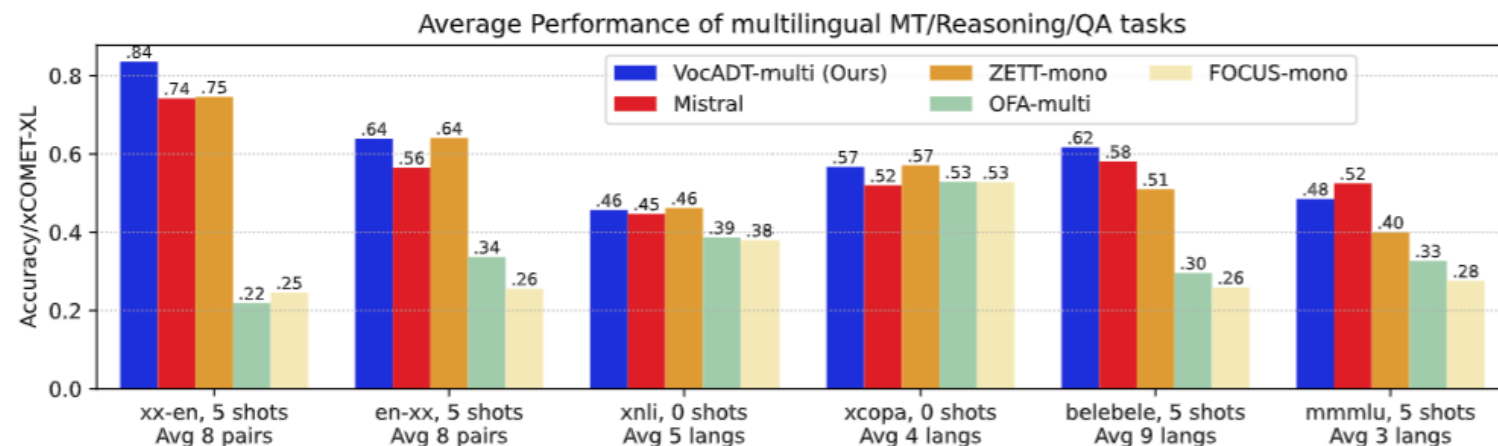
	LLaMA2							
	MT		XNLI		QA		XLSUM	
	En-X	X-En	en	avg	en	avg	en	avg
CW2V	17.0	27.3	60.4	38.1	77.7	35.8	0.6	0.4
OFA	11.2	16.2	60.4	37.1	76.0	26.0	0.6	0.3

CW2V is Word2Vec objective

Approaches which learn the weights outperform or are at least competitive with approaches using heuristics

Simpler training pipelines, require less external resources

Note: ZETT also uses LM objective and a more general mapping function



VocADT multi is the LM objective approach

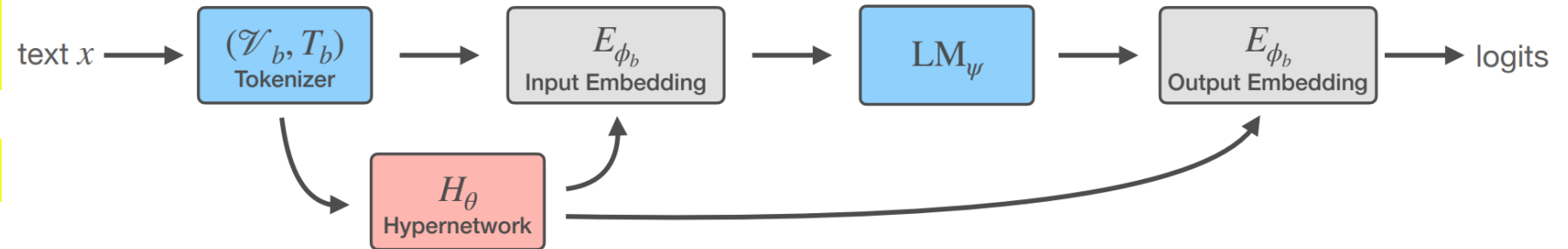
Zero-shot Tokenizer Transfer [ZETT]

Learn a function to predict embeddings for any given tokenizer for a fixed language model?

Train once, and then use it to predict embeddings for any new tokenizer

Hypernetwork generates target embeddings for the new tokenizer

Replaces static embedding lookup



Training the hypernetwork

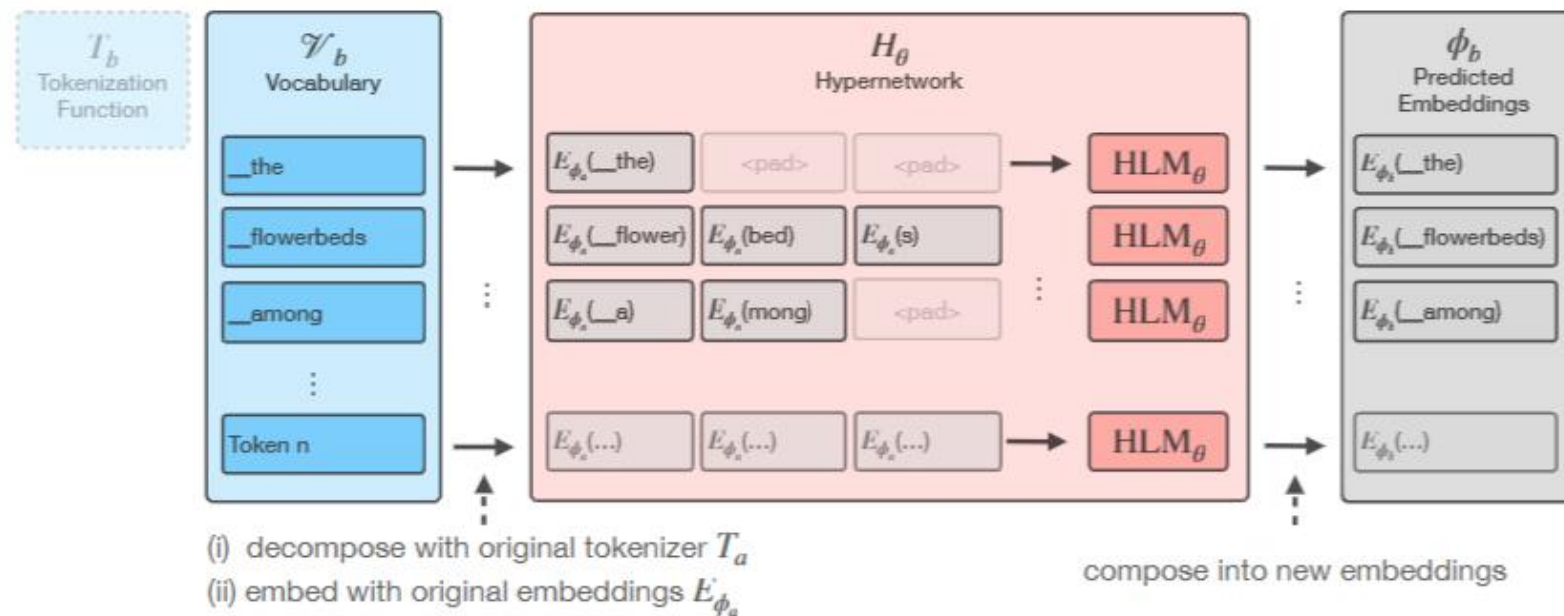
Diverse Tokenizer data

- Sample frequent tokens from all substrings
- Sampled from different data subsets
- Noise added to frequencies

Training Objective (End-to-End training)

- Gold Target embeddings are not available
- Train with standard LM objective (MLM, CLM)
- Only the Hypernetwork parameters are trained
- Main LM parameters are frozen

Hypernetwork design [ZETT]



HLM is a small encoder LM that generates embeddings for a new token from a sequence of original tokens

Hypernetwork is trained to minimize the loss of the main LLM

Making hypernetwork training easier – First train to mimic embeddings of the initial tokenizer

Stabilize training: auxiliary loss to minimize difference between new and original embeddings

Results for ZeTT [ZETT]

ZETT performs better than other approaches on multiple downstream tasks

	ar	bg	de	el	en	es	fr	hi	ru	sw	tr	ur	vi	Avg.
original	68.9	75.6	74.7	73.7	82.3	76.9	76.8	68.4	72.9	63.5	72.2	64.7	73.1	72.6
Lexical	58.7	63.1	65.3	61.7	72.8	68.4	66.7	61.8	62.3	51.8	58.5	60.0	72.0	63.3
FVT	63.9	70.3	70.9	67.4	79.0	73.9	71.9	65.7	67.8	57.1	66.3	61.7	72.9	68.4
OFA	57.3	64.2	67.3	62.8	73.6	68.6	68.4	61.8	63.1	54.8	59.7	59.3	72.3	64.1
FOCUS	64.8	71.0	71.6	67.7	79.6	74.4	72.6	64.5	68.1	55.7	67.3	61.9	72.6	68.6
ours	67.9	73.9	74.1	71.4	81.1	76.2	74.7	67.7	70.7	62.3	68.7	63.2	73.9	71.2
Δ accuracy	-1%	-2%	-1%	-2%	-1%	-1%	-2%	-1%	-2%	-1%	-3%	-2%	+1%	-1%
Δ length	-22%	-14%	-13%	-23%	-9%	-11%	-12%	-13%	-13%	-19%	-15%	-9%	-3%	-14%

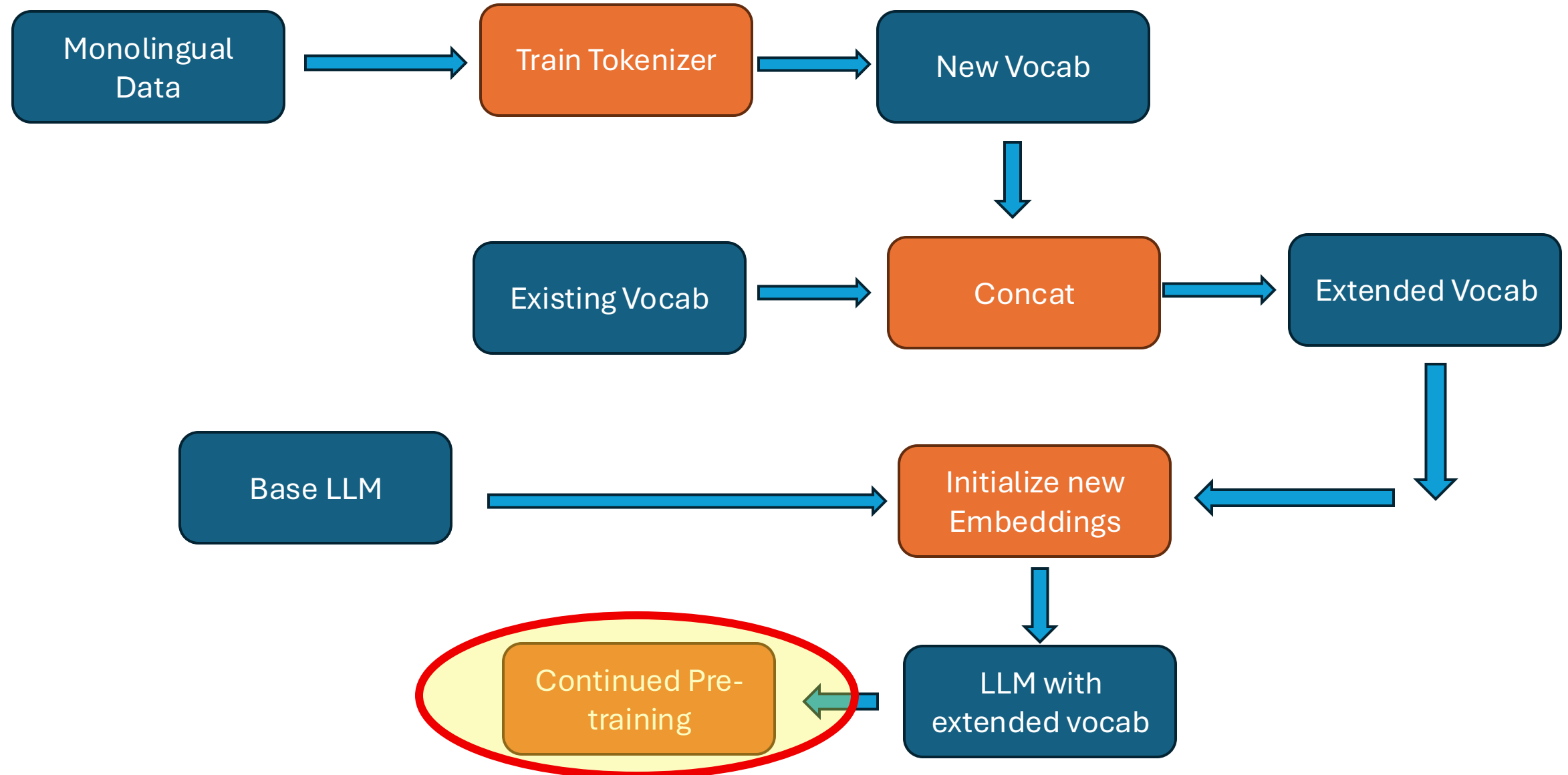
CPT with less than a billion tokens can close the gap with the original model

Hypernetwork trained for base model can be applied to fine-tuned model

#shots Method	Natural Language ($\rightarrow$ GPT2 Tok.)					
	PiQA	HS	ARC	BoolQ	MMLU	Avg.
original	80.7	81.0	79.5	83.6	59.6	76.9
original@800M	82.1	82.7	80.6	80.6	57.8	76.8
0-shot FOCUS	69.2	63.8	45.7	60.4	38.8	55.6
0-shot ours	79.7	77.5	73.0	81.9	53.0	73.0
n -shot FOCUS@800M	74.8	74.3	72.4	73.3	48.9	68.7
n -shot ours@800M	80.9	80.7	77.8	80.7	54.4	74.9

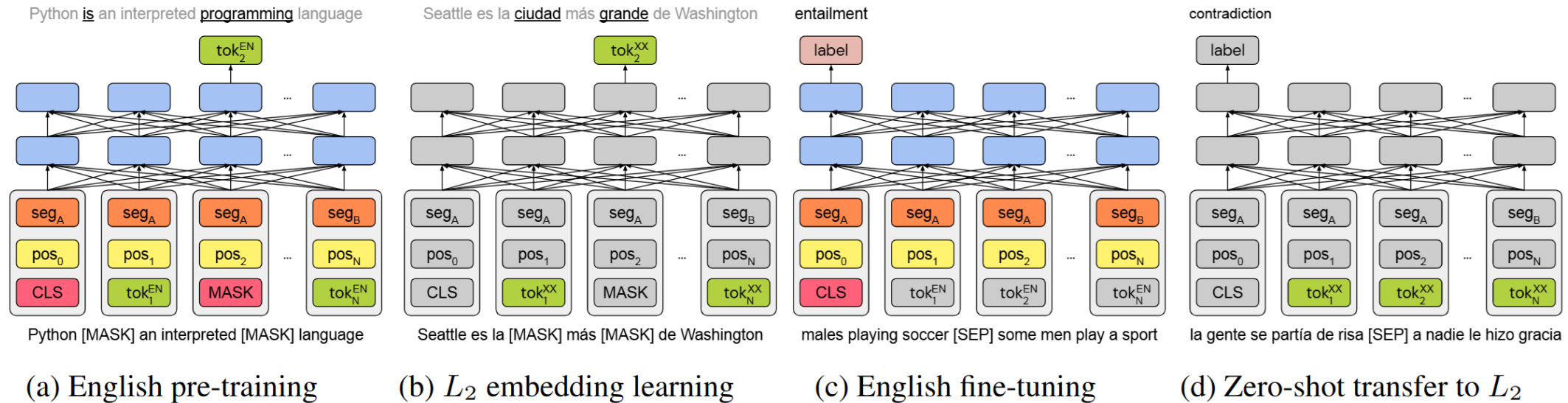
Embeddings	original		0-shot		n-shot			
	orig.	base	FOCUS	ours	ours@800			
λ	-	-	-	-	0.0	0.3	0.5	0.7
Score (1 to 10)	7.33	7.48	5.03	6.56	6.59	6.75	6.82	6.77

How to extend tokenizer vocabulary?



Basic CPT Recipe for vocab transfer

[CLT]



CPT: train only embeddings

Task-training: train only non-embedding parameters

Other CPT choices include

Language-specific positional embeddings

Adding noise to embeddings

Language-specific adapters

Results

Method	XNLI	MLDOC	PAWS-X	XQuAD
Joint Multilingual	70.5	80.4	87.0	65.7
Joint Pair	71.1	80.8	87.5	67.0
Adapted Model				
vanilla	67.8	81.2	86.0	61
+ pos emb	69.1	80.0	86.9	56.8
+ noising	70.0	76.7	82.7	57.8
+ adapters	69.5	79.3	82	66.8

(average results across multiple languages)

- **Adapted model is competitive with the jointly trained models**
- Different design choices are useful in different scenarios (language-specific variations)
 - **Language-specific positional embeddings are generally useful**
- **English performance is not degraded**
- **Generalization across languages in the internal representations of the network**

Some CPT strategies for faster convergence

- Select high new-token coverage data *[DataSelect]*
- Initially freeze non-embedding parameters *[CLT]*
- For shared vocab items – try not to deviate much from original embeddings *[ZETT]*
- Learn new embeddings using a pruned LLM when using end2end LM objective aka VocADAPT and ZeTT *[MiniModel]*

Can initial pre-training be designed so that it is easier to expand vocabulary and support new languages later?

Goal: Improve plasticity of initial pre-trained model

Universal Tokenizer *[UniversalTok]*

Train tokenizer to support more languages than supported during initial pre-training

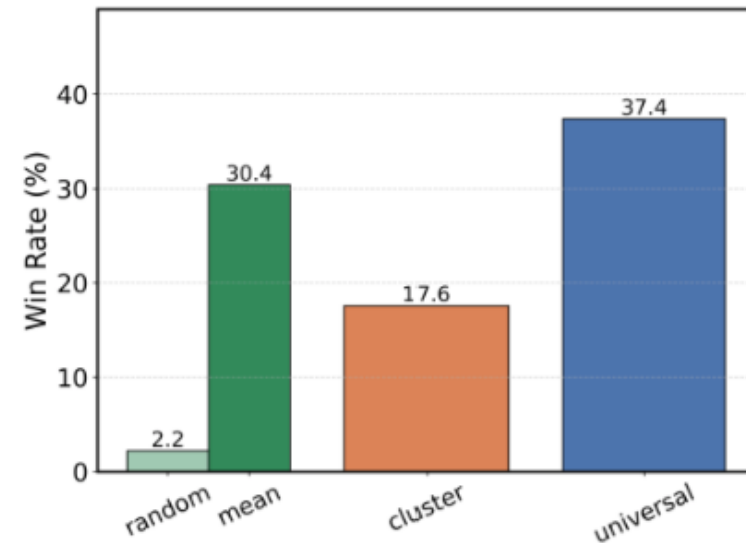
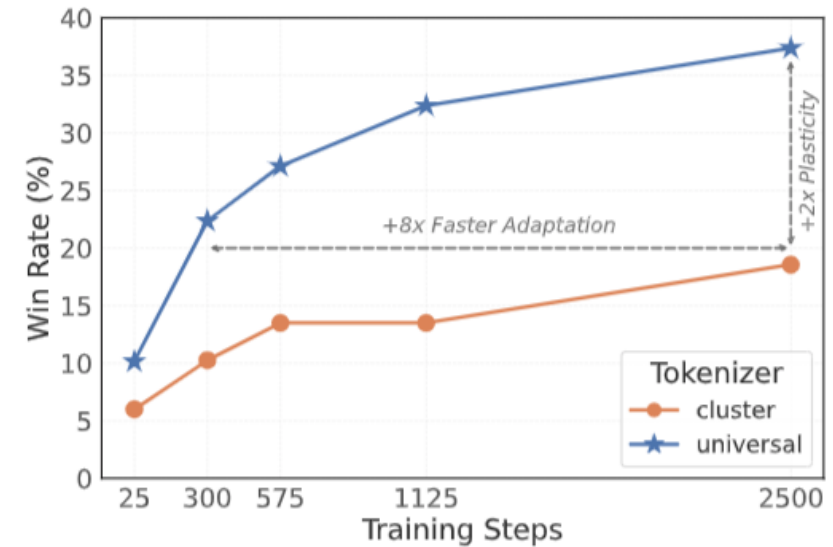
CPT benefits more when using a tokenizer that supports expanded languages.

Even languages not included in universal tokenizer benefit better during CPT

Improved language plasticity and faster adaptation

Can outperform strong vocabulary initialization methods like average of source vocab embeddings

Works probably because of data leakage in the initial pre-training

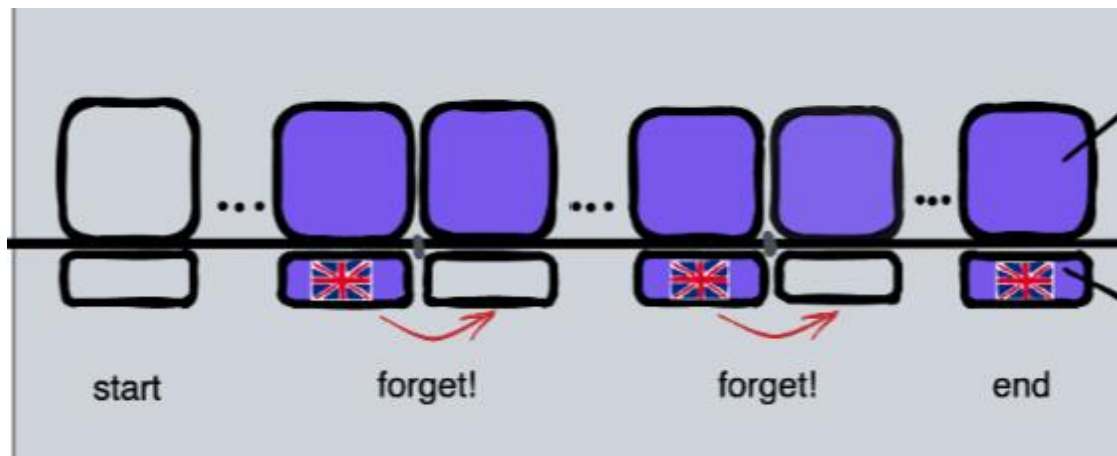


Pre-training with active forgetting

[ActiveForget]

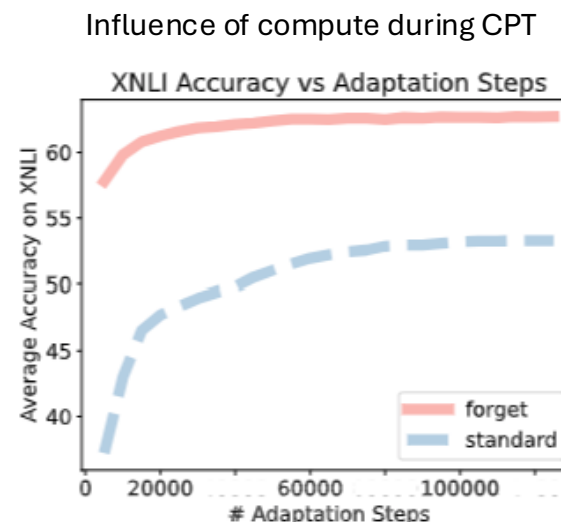
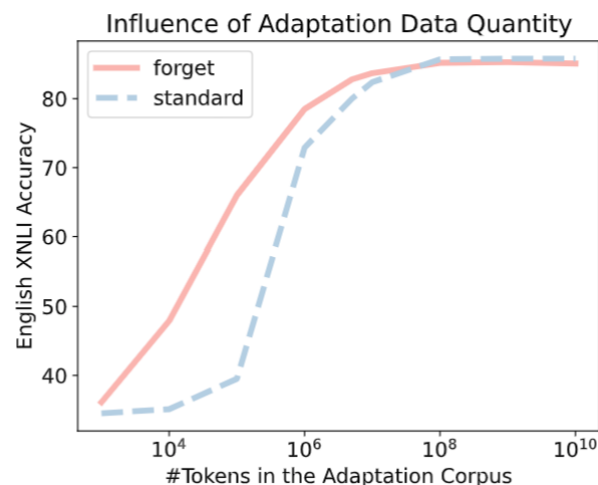
During initial pre-training, reset the embeddings periodically

Avoids network taking shortcuts in learning embeddings



Teaches the network to adapt embeddings quickly during continued pre-training

Forgetting during network training has also been explored in more general settings



Summary of initialization methods

Sampling from Random (Normal) Distribution

Simple

Changes existing vocab's probability distribution

Large convergence time

Average of Existing Embeddings

Limited change in existing vocab's distribution

Large convergence time [AveInit]

Weighted Average of Existing Embeddings

Initializations like [WECHSEL, OFA, FOCUS, ConstrainedW2V, VocAdapt]

Improves convergence rates

Hypernetworks for learning embeddings

Learn a hypernetwork that can predict embeddings for any tokenizer, enabling zero-shot tokenizer transfer [ZETT]

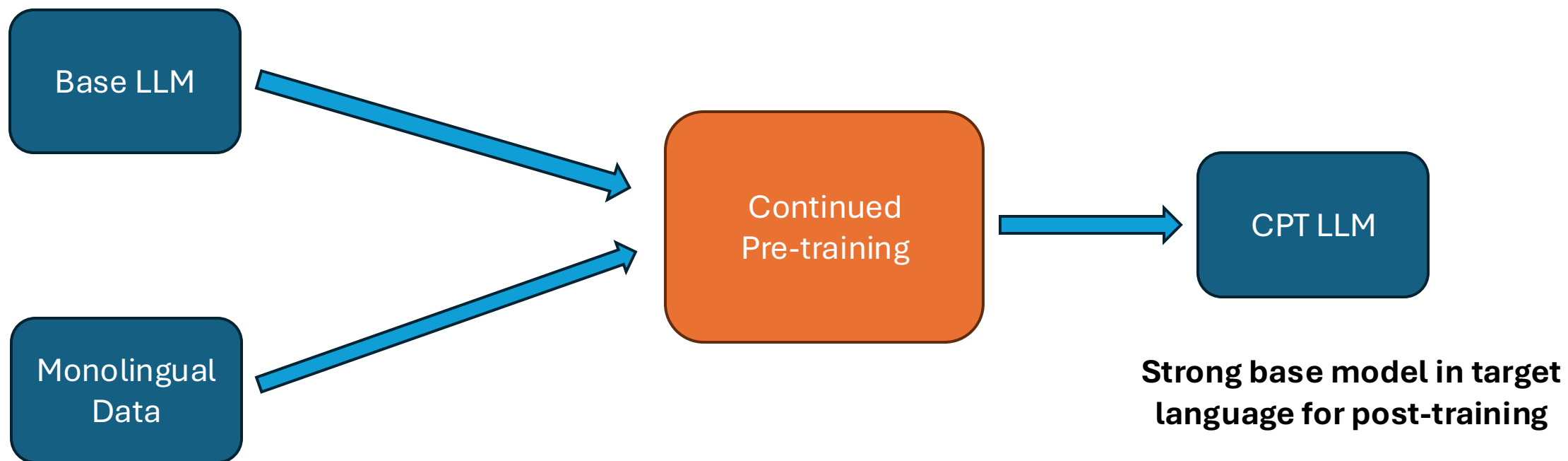
Summary & Recommendations

- **Vocab expansion reduces fertility and improves efficiency**
- **Can we do better than random initialization?**
 - Weighted averaging and methods learning the new embeddings do better
 - Simple methods like averaging, constrained W2V are strong baselines
 - Fast convergence with less than a billion tokens
- **Will vocabulary expansion lead to lower performance on English?**
 - If initialized embeddings are in convex hull, greedy decoding results does not change
 - CPT can lead to initial drop in results – can be avoided or mitigated with the right representation of English in data mixture, CPT method, etc.
- **Is vocabulary expansion competitive/better than relying to initial sub-optimal vocab?**
 - Yes, with the right choices about embedding initialization, CPT, amount of data, etc.
- **Make provision for vocabulary expansion in initial pre-training**

Tutorial Outline

- Introduction
 - The Multilingual LLM Challenge
- Extending LLMs to new languages
 - Vocabulary Expansion
 - **Continued Pre-training**
 - Post-Training
- Multilingual Evaluation
- Summary & Future Directions

Continued Pre-training



Causal Language Modeling Objective

$$p(\mathbf{x}) = p(x_1, x_2, \dots, x_T) = \prod_{t=1}^T p(x_t | \mathbf{x}_{<t})$$

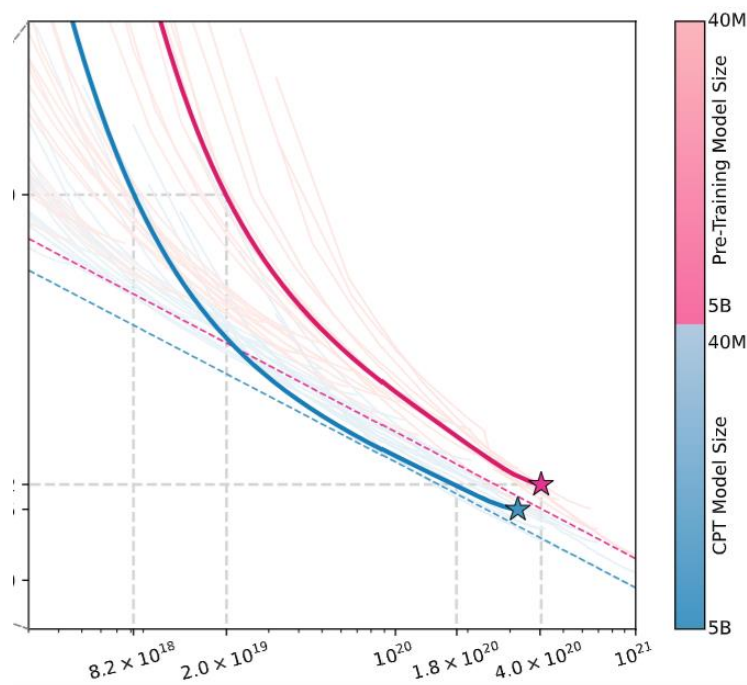
Why do continued pre-training? (1)

Language competence/fluency in target language

Perplexity reduces with increase in pre-training corpus size

	L(0)	L(10k)	L(100k)	L(1M)
Chinese	10.151	8.697	6.634	5.249

[LmaByndEng]



Achieves same loss as full pre-training from scratch for new language with far less compute – 25 to 50% lower

[ScalingCPT]

Why do continued pre-training? (2)

Improve alignment b/w English and target language

Language	Base LLM	After CPT
Gujarati	0.39	0.46
Hindi	0.40	0.44
Marathi	0.44	0.48

Cosine similarities between English and target languages increases with CPT

[RomanSetu]

This can be useful for transfer learning from English

Why do continued pre-training? (3)

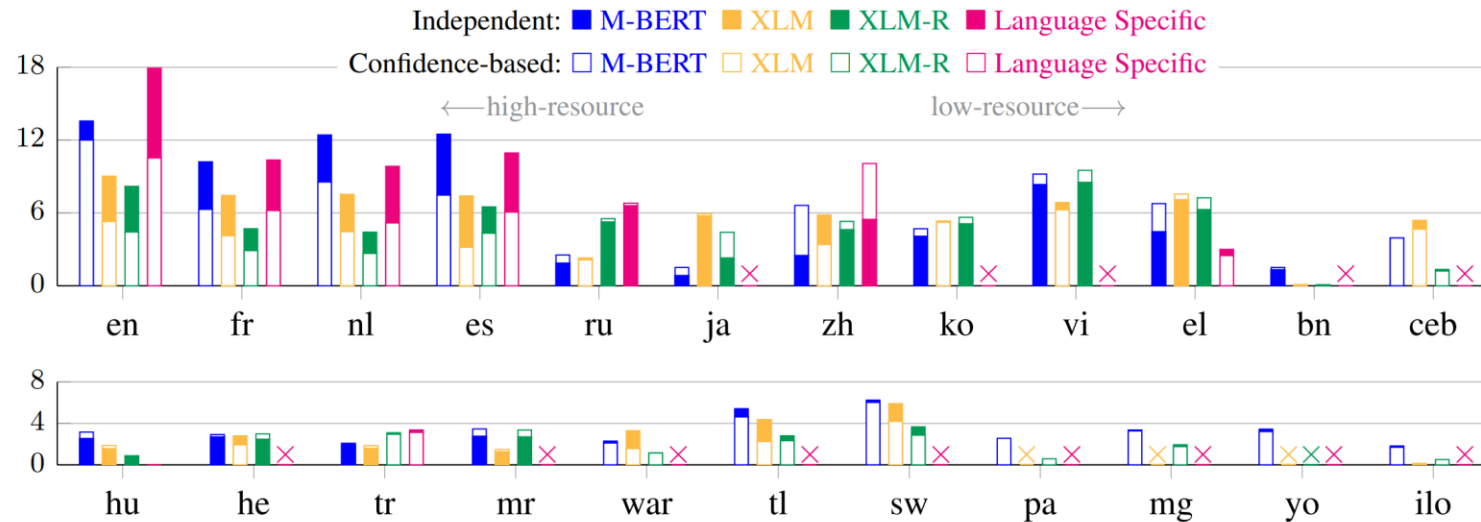
Provide required knowledge in target language for better understanding

- **LMs better at using in-language knowledge vs. cross-lingual transfer**
- Incorporate cultural-specific knowledge captured in native language corpora only

Knowledge Probing Task: Predict missing tokens which capture model's knowledge

fact ⟨Bloomberg L.P., founded_in, New York⟩
en prompt [X] was founded in [Y].

es sentence	Bloomberg L.P. fue fundada en ⟨mask⟩ × 1 ~ 5.		
	prediction	#tokens	confidence
es outputs	2012	1	-1.90
	Nueva York	2	-0.61
	EE. UU	3	-1.82
	Chicago, Estados Unidos	4	-3.58
	2012 Bloomberg L.P	5	-3.06



English accuracy for knowledge probing is much higher than other languages

Results on Knowledge Probing task shows that non-English languages don't have enough data

Most multilingual models can't transfer knowledge in English to other languages

Why do continued pre-training? (4)

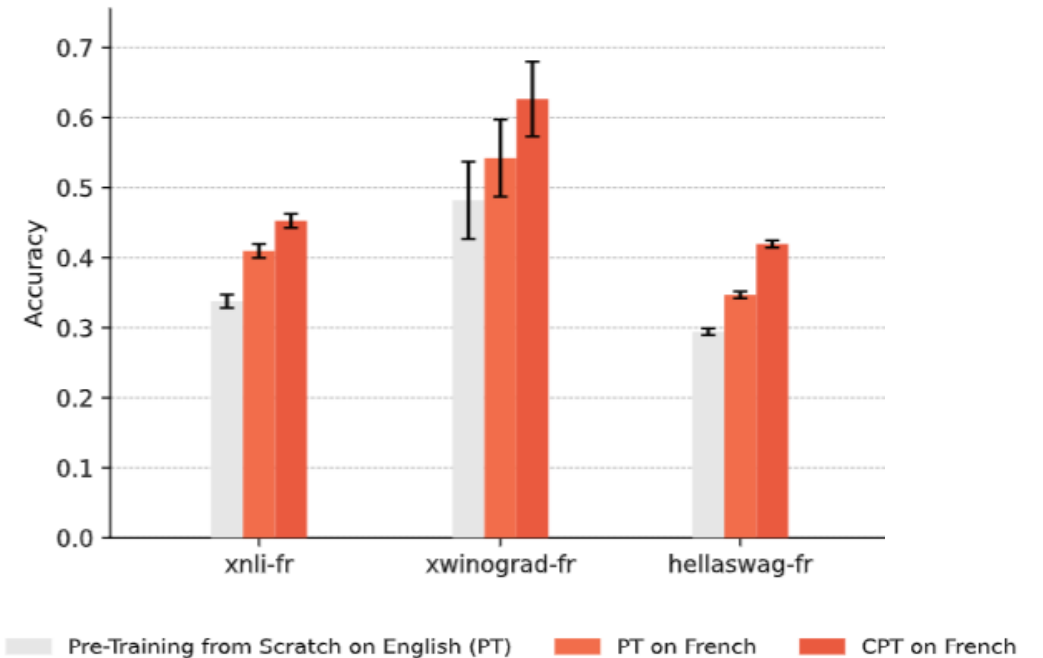
Improved downstream task performance

	PPL	Dwn.	Cop.
Llama-2 (7B)	23.64	27.43	44.67
+ CPT (eu+en)	3.35	34.14	43.43

[EmergentCPT]

CPT can help improve

- Language Fluency
- Downstream Task performance



Retain in-context capabilities of the original model

[ScalingCPT]

Better than pre-training from scratch for a new language

[RomanSetu, EmergentCPT, ScalingCPT]

***What properties do
we want to see in
multilingual corpora?***

Large-scale, Document-level Datasets

High Quality Documents

Wide coverage of topics

Representation of culture-specific data, native literature

Capture data in different modalities and genres

Data to Help Cross-lingual transfer with English

Publicly Multilingual corpora are good starting points



Large-scale, Document-level Datasets



High Quality Documents



Wide coverage of topics



Representation of culture-specific data, native literature



Capture data in different modalities and genres



Data to Help Cross-lingual transfer with English

Major Corpora

mC4, CC100

Wikipedia

OSCAR

CulturaX

MADLAD

Glott-500

MALA-500

Build custom language (group) specific collections to address gaps

FineWeb2 – Scaling up the Web Data

- ~100 common crawl snapshots
- 1000+ languages
- 20TB of data
- 5B documents
- 3T words

*But web-data
independently may not
be sufficient...*

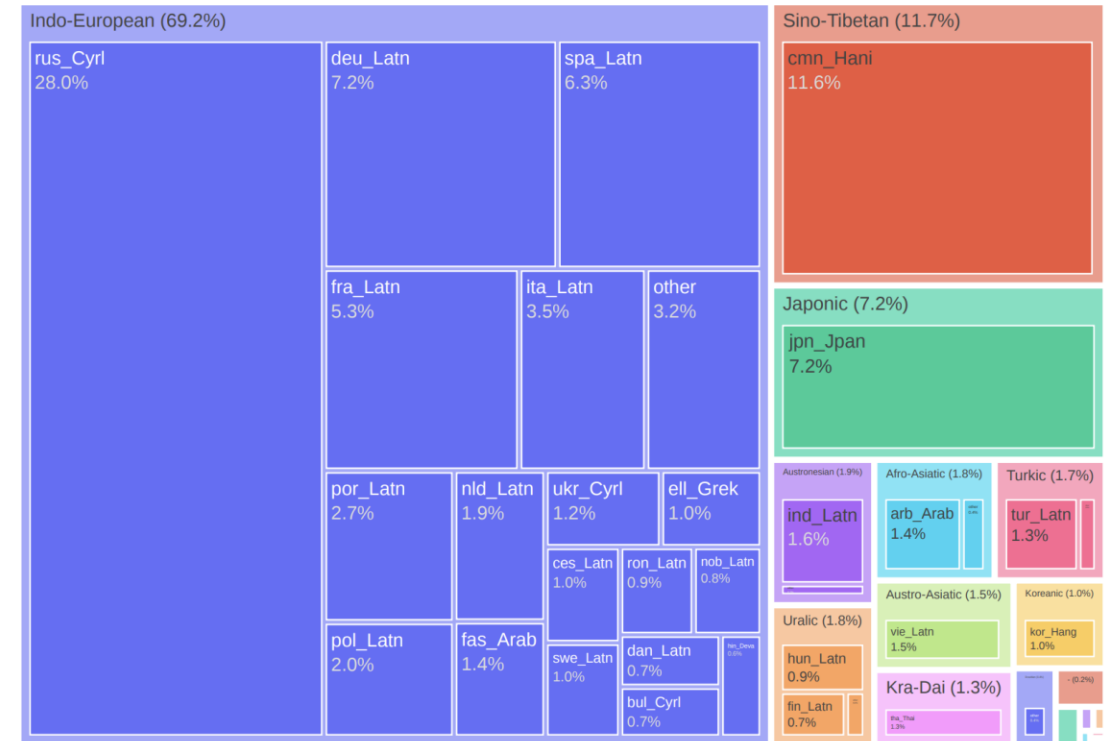


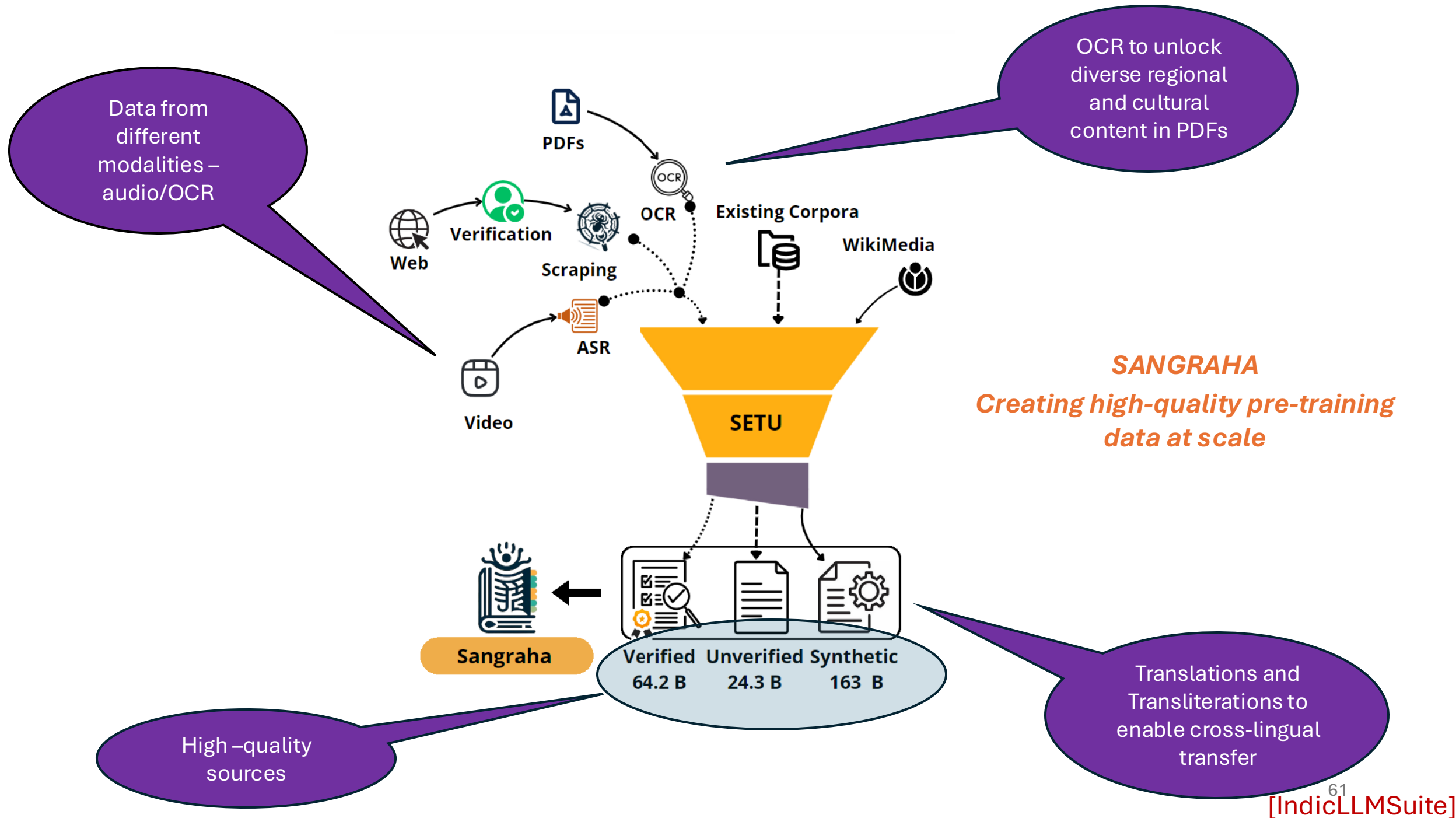
Figure 8: **Language composition of FineWeb2** Distribution of languages in the final FineWeb2 dataset. Percentages refer to total utf-8 bytes of each language or language family.

[FineWeb2]

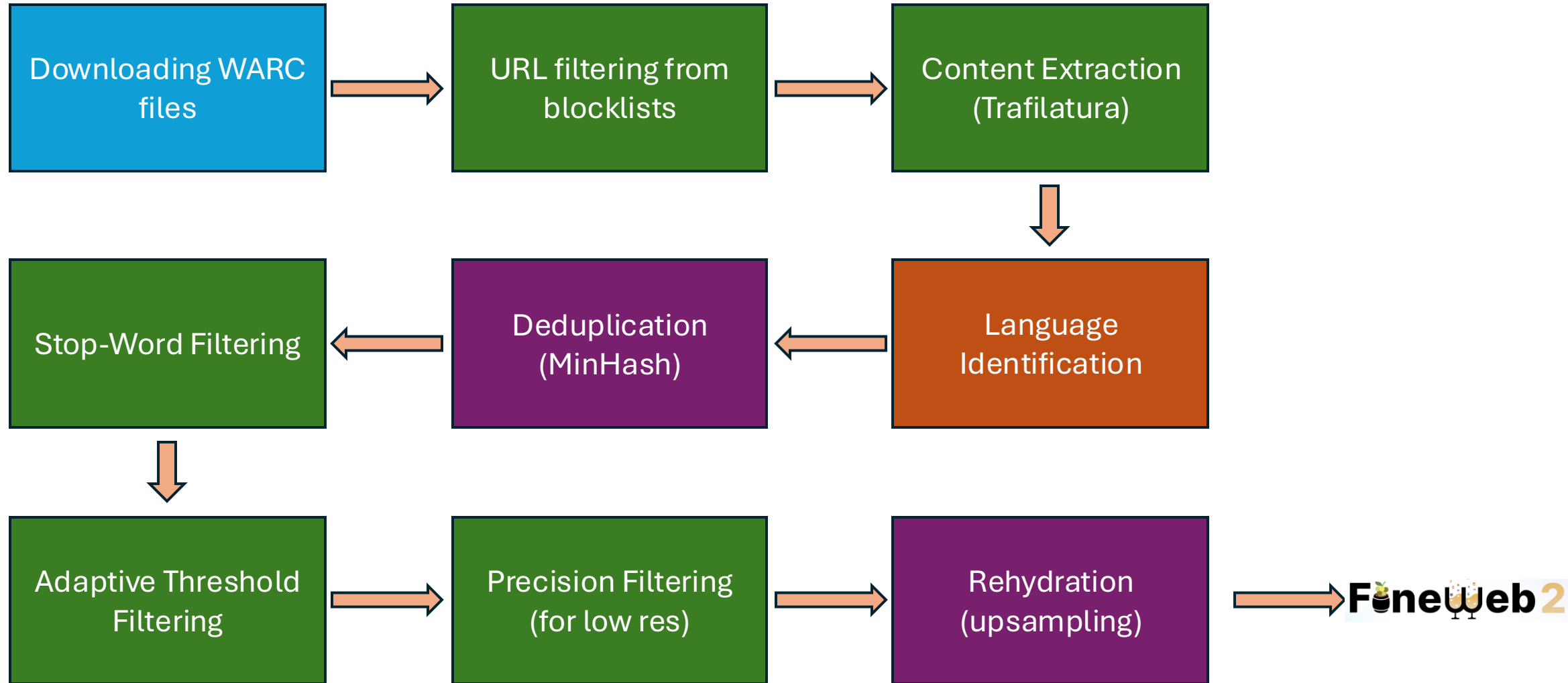
ISO 639-3 code	Script	Name	Language Family	Subset	Words	Documents	UTF-8 Bytes	Disk size
rus	Cyrl	Russian	Indo-European	rus_Cyrl	588,579,493,780	699,083,579	5.82TB	1.81TB
cmn	Hani	Mandarin Chinese	Sino-Tibetan	cmn_Hani	543,543,038,750	636,058,984	2.42TB	1.48TB
deu	Latn	German	Indo-European	deu_Latn	262,271,052,199	495,964,485	1.51TB	719.08GB
jpn	Jpan	Japanese	Japonic	jpn_Jpan	331,144,301,801	400,138,563	1.50TB	667.44GB
spa	Latn	Spanish	Indo-European	spa_Latn	261,523,749,595	441,287,261	1.32TB	593.82GB
fra	Latn	French	Indo-European	fra_Latn	220,662,584,640	360,058,973	1.11TB	502.82GB
ita	Latn	Italian	Indo-European	ita_Latn	139,116,026,491	238,984,437	739.24GB	332.47GB
por	Latn	Portuguese	Indo-European	por_Latn	109,536,087,117	199,737,979	569.24GB	256.92GB
pol	Latn	Polish	Indo-European	pol_Latn	73,119,437,217	151,966,724	432.01GB	210.35GB
hin	Deva	Hindi	Indo-European	hin_Deva	11,173,681,651	22,095,985	120.98GB	31.92GB
ben	Beng	Bengali	Indo-European	ben_Beng	6,153,579,265	15,185,742	87.04GB	22.25GB
slk	Latn	Slovak	Indo-European	slk_Latn	14,808,010,769	29,991,521	85.43GB	43.00GB
heb	Hebr	Hebrew	Afro-Asiatic	heb_Hebr	8,462,976,117	14,491,748	68.71GB	23.15GB
lit	Latn	Lithuanian	Indo-European	lit_Latn	9,132,828,961	13,471,965	56.50GB	25.75GB
bos	Latn	Bosnian	Indo-European	bos_Latn	9,086,837,979	21,243,255	49.18GB	24.61GB
slv	Latn	Slovenian	Indo-European	slv_Latn	7,688,373,264	12,059,130	41.80GB	19.22GB

Explore additional sources of data..

[FineWeb2]



Pipelines for Cleaning Web Data



Examples of Filtering Thresholds

Metrics	Description
<i>bytes</i>	size of the document interms of bytes,
<i>word_count</i>	no.of words present in a document
<i>char_count</i>	no.of characters present in a document
<i>lines_count</i>	total no.of sentences present in a document
<i>mean_line_length</i>	mean sentence length interms of words of a document
<i>min_line_length</i>	minimum sentence length interms of words of a document.
<i>max_line_length</i>	max sentence length interms of words of a document
<i>nsfw_words_count</i>	no.of NSFW words present in a document
<i>non_li_character_count</i>	no.of non-latin/non-indic characters in a document
<i>10_gram_characters_repetition_score</i>	score used for filtering documents using 10-gram character repetition filter
<i>5_gram_words_repetition_score</i>	score used for filtering documents using 5-gram word repetition filter

Table 6: Showing all the metrics that are calculated in analysis stage

**Common
VS
Language Specific
thresholds?**

Adaptive Filtering Thresholds

Filter Group			cc				wiki			
	Baseline	English	10Tail	MeanStd	MedRatio	Quant	10Tail	MeanStd	MedRatio	Quant
fwq	7.00	-	-	5.22	4.00	4.33	3.00	5.00	3.89	3.56
goq	6.33	-	5.22	-	3.89	4.56	4.44	4.11	4.22	3.22
gor	6.22	4.22	3.33	2.22	-	4.11	-	3.89	-	4.00

Table 25: **Average ranks** by block and method across all languages. *Baseline* has no filtering, English is the default FineWeb English thresholds. We then compute each of the other 4 methods – 10Tail, MeanStd, MedRatio (MedianRatio), and Quantile (Quant) – on both Common Crawl (cc) data and on Wikipedia (wiki). Cells marked with - correspond to method-filter-group combinations that would remove over 75% of data with a single filter on at least one of the languages, or that would not remove anything at all. Lower ranks are better.

Multilingual Challenges in Filtering

URL filtering from
blocklists

Can existing blocklists or heuristics be extended to multilingual content?

Stop-Word Filtering

Need to curate language specific stop-word lists or toxic-word lists

Language
Identification

Are existing LID models good enough for code-mixed, Romanized content and distinguishing similar languages?

Examples of Filtered Content

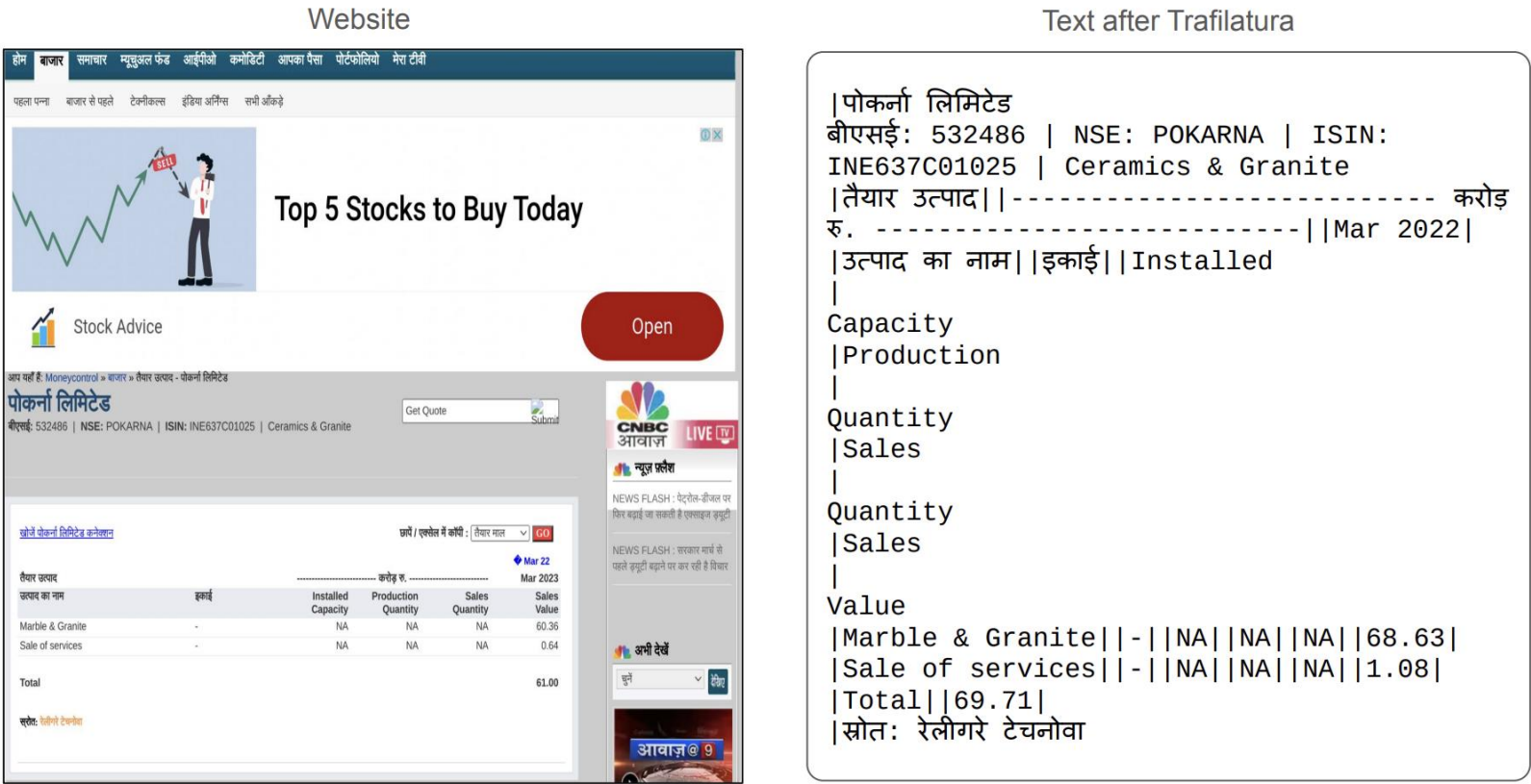


Figure 6: Document flagged by symbol heavy filter in the Cleaning and Analysis stage.

Examples of Filtered Content

Uncleaned Text

MaxCyte Signs Strategic Platform License with Lyell

Lyell Immunopharma to use MaxCyte's Flow Electroporation:registered: technology and ExPERT:tm: platform in its T cell product candidates targeting solid tumors.

ROCKVILLE, Md., July 06, 2023 (GLOBE NEWSWIRE) – MaxCyte, Inc., (Nasdaq: MXCT; LSE: MXCT), a leading, cell-engineering focused company providing enabling platform technologies to advance the discovery, development and commercialization of next-generation cell-based therapeutics and to support innovative, cell-based research, today announced the signing of a strategic platform license (SPL) with Lyell Immunopharma, Inc., a clinical stage T cell reprogramming company.

Under the terms of the agreement, Lyell Immunopharma obtains non-exclusive clinical and commercial rights to use MaxCyte's Flow Electroporation:registered: technology and ExPERT:tm: platform. In return, MaxCyte is eligible to receive platform licensing fees, clinical milestone payments and sales-based payments.

"At MaxCyte, our goal is to maximize the potential of cells to improve patients' lives, and it is through collaborations such as this that we can achieve success," said Doug Doerfler, President and CEO of MaxCyte. "We look forward to supporting Lyell Immunopharma in its development of solid tumor treatments for patients with unmet needs." MaxCyte's ExPERT:tm: instrument portfolio is the next generation of leading, clinically-validated electroporation technology for complex and scalable cell engineering. By delivering high transfection efficiency, seamless scalability and enhanced functionality, the ExPERT:tm: platform delivers the high-end performance essential to enabling the next wave of biological and cellular therapeutics. Each of MaxCyte's strategic partnerships generates pre-commercial milestone revenue and the vast majority include sales-based payments.

About MaxCyte

At MaxCyte, we pursue cell engineering excellence to maximize the potential of cells to improve patients' lives. We have spent more than 20 years honing our expertise by building best-in-class platforms, perfecting the art of the transfection workflow, and venturing beyond today's processes to innovate tomorrow's solutions. Our ExPERT:tm: platform, which is based on our Flow Electroporation:registered: technology, has been designed to support the rapidly expanding cell therapy market and can be utilized across the continuum of the high-growth cell therapy sector, from discovery and development through commercialization of next-generation, cell-based medicines. The ExPERT family of products includes: four instruments, the ATx:tm:, STx:tm:, GTx:tm: and VLx :tm:; a portfolio of proprietary related processing assemblies or disposables; and software protocols, all supported by a robust worldwide intellectual property portfolio. By providing our partners with the right technology, as well as technical and regulatory support, we aim to guide them on their journey to transform human health. Learn more at maxcyte.com and follow us on Twitter and LinkedIn.

MaxCyte Contacts:

US IR Adviser
Gilmartin Group
David Deuchler, CFA
+1 415-937-5400
ir@maxcyte.com

US Media Relations
Spectrum Seismic Collaborative
Valerie Enes
+1 408-497-8568
valerie@spectrumscience.com

Nominated Adviser and Joint Corporate Broker
Panmure Gordon
Emma Earl / Freddy Crossley
Corporate Broking
Rupert Dearden
+44 (0)20 7886 2500

UK IR Adviser
Consilium Strategic Communications
Mary-Jane Elliott
Chris Welsh
+44 (0)203 709 5700
maxcyte@consilium-comms.com

Clean Text after Stage 2

Lyell Immunopharma to use MaxCyte's Flow Electroporation:registered: technology and ExPERT:tm: platform in its T cell product candidates targeting solid tumors.

ROCKVILLE, Md., July 06, 2023 (GLOBE NEWSWIRE) – MaxCyte, Inc., (Nasdaq: MXCT; LSE: MXCT), a leading, cell-engineering focused company providing enabling platform technologies to advance the discovery, development and commercialization of next-generation cell-based therapeutics and to support innovative, cell-based research, today announced the signing of a strategic platform license (SPL) with Lyell Immunopharma, Inc., a clinical stage T cell reprogramming company.

Under the terms of the agreement, Lyell Immunopharma obtains non-exclusive clinical and commercial rights to use MaxCyte's Flow Electroporation:registered: technology and ExPERT:tm: platform. In return, MaxCyte is eligible to receive platform licensing fees, clinical milestone payments and sales-based payments.

"At MaxCyte, our goal is to maximize the potential of cells to improve patients' lives, and it is through collaborations such as this that we can achieve success," said Doug Doerfler, President and CEO of MaxCyte. "We look forward to supporting Lyell Immunopharma in its development of solid tumor treatments for patients with unmet needs."

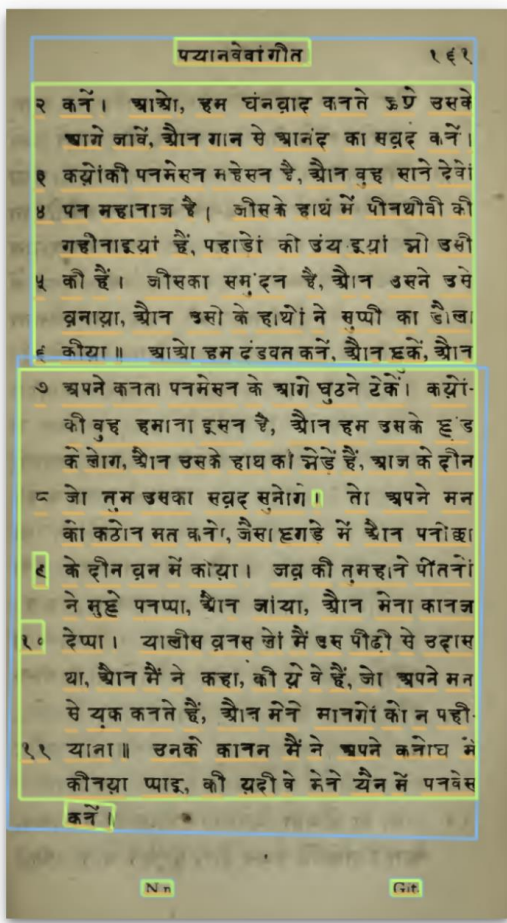
MaxCyte's ExPERT:tm: instrument portfolio is the next generation of leading, clinically-validated electroporation technology for complex and scalable cell engineering. By delivering high transfection efficiency, seamless scalability and enhanced functionality, the ExPERT:tm: platform delivers the high-end performance essential to enabling the next wave of biological and cellular therapeutics. Each of MaxCyte's strategic partnerships generates pre-commercial milestone revenue and the vast majority include sales-based payments.

At MaxCyte, we pursue cell engineering excellence to maximize the potential of cells to improve patients' lives. We have spent more than 20 years honing our expertise by building best-in-class platforms, perfecting the art of the transfection workflow, and venturing beyond today's processes to innovate tomorrow's solutions. Our ExPERT:tm: platform, which is based on our Flow Electroporation:registered: technology, has been designed to support the rapidly expanding cell therapy market and can be utilized across the continuum of the high-growth cell therapy sector, from discovery and development through commercialization of next-generation, cell-based medicines. The ExPERT family of products includes: four instruments, the ATx:tm:, STx:tm:, GTx:tm: and VLx :tm:; a portfolio of proprietary related processing assemblies or disposables; and software protocols, all supported by a robust worldwide intellectual property portfolio. By providing our partners with the right technology, as well as technical and regulatory support, we aim to guide them on their journey to transform human health. Learn more at maxcyte.com and follow us on Twitter and LinkedIn.

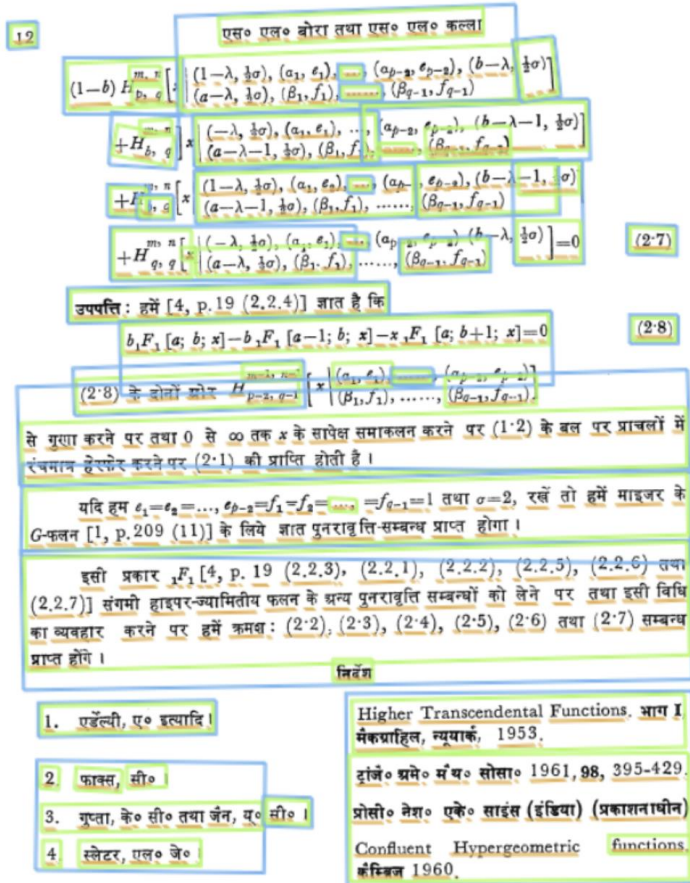
MaxCyte Contacts:

Figure 7: Cleaning performed by 'terminal punctuation filter' in Cleaning and Analysis stage.

Multimodal Filters



(a) **Bounding Box Suppression:** Page in which smaller bounding boxes are suppressed as these can lead to false flagging of pages or misaligned text.



(d) **High Bounding Box Overlap:** Page filtered out due to high bounding box overlap. This high overlapping can lead to disordered parsing of text, break in continuity, etc.

[IndicLLMSuite]

Multimodal Filters

amendment	संशोधन
amenity	सुख-सुविधा
amicable	सौहार्दपूर्ण, मैत्रीपूर्ण
amnesty	सर्वश्रमा, सामाहिक क्षमादान
analyst	विश्लेषक
anarchist	१. अराजकतावादी २. आतंकवादी
anarchy	अराजकता
animated suspension	सक्रिय निवर्तन
annex	१. संलग्न करना, नत्थी करना २. राज्य में मिला लेना
annexation	समामेलन, राज्य में मिला लेना
annexe	उपभवन, एनेक्सी, सौध
annexed territory	समाभिलिप्त भूभाग
annexure	संलग्नक
announcement	आख्यापन, ऐलान, उद्घोषणा
annual maintenance contract	वार्षिक रख-रखाव संधि
annual report	वार्षिक रिपोर्ट
anomaly	असंगति
anonymous complaint	गुमनाम शिकायत, अनाम शिकायत
ante chamber	रजकक्ष
anticipation	१. प्रत्याशा २. पूर्वनिर्माण
anti-national activities	राष्ट्रविरोधी गतिविधियाँ
antinationalism	राष्ट्रवाद विरोध
anywhere in India	भारत में कहीं भी
apex	शीर्ष
apology	क्षमायाचना
appeal	टपील, अपील करना
appeasement	तृप्तीकरण

7

(b) **Horizontally Sparse:** Page filtered out due to less horizontal text coverage, this can be indicative of very small lines, lists, index etc.

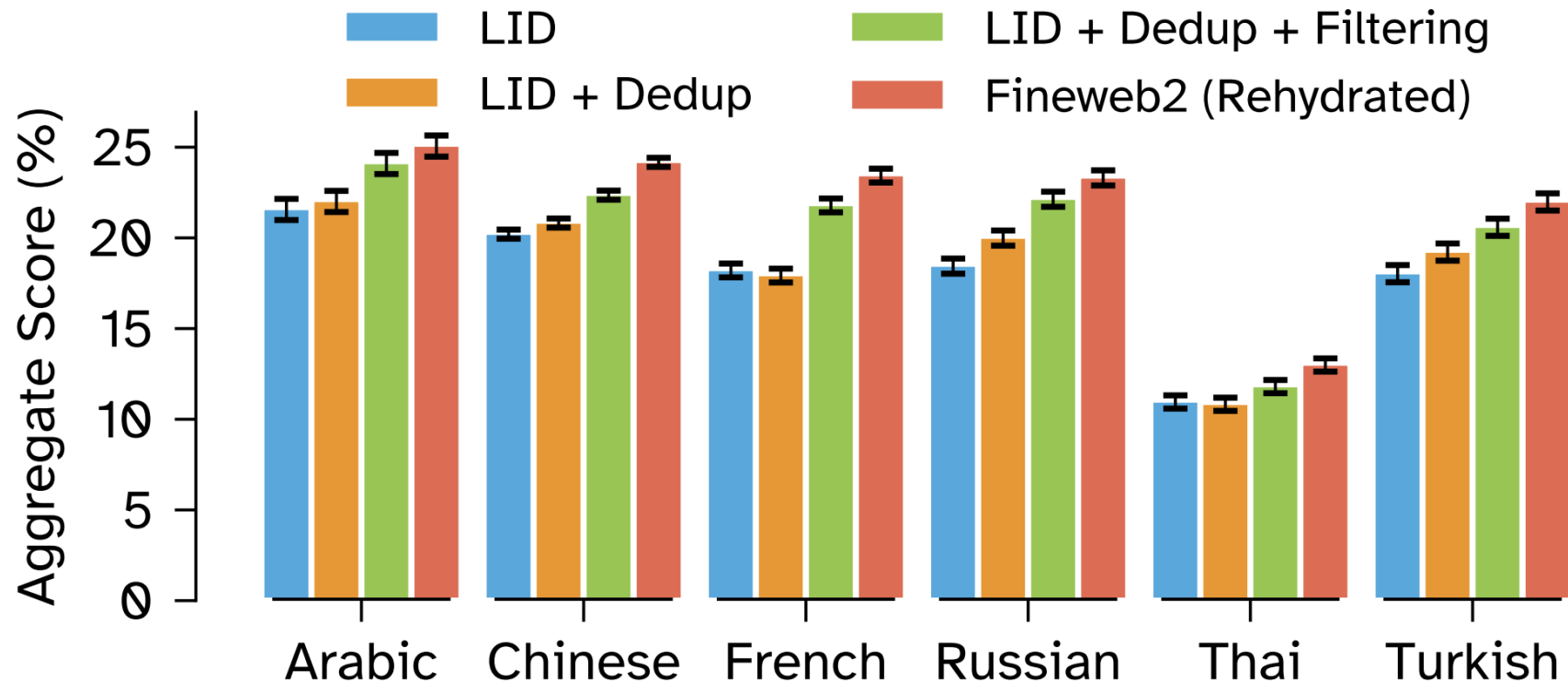


2

(c) **Vertically Sparse:** Page filtered out due to less vertical text coverage. This can be indicative of title pages, comics, etc.

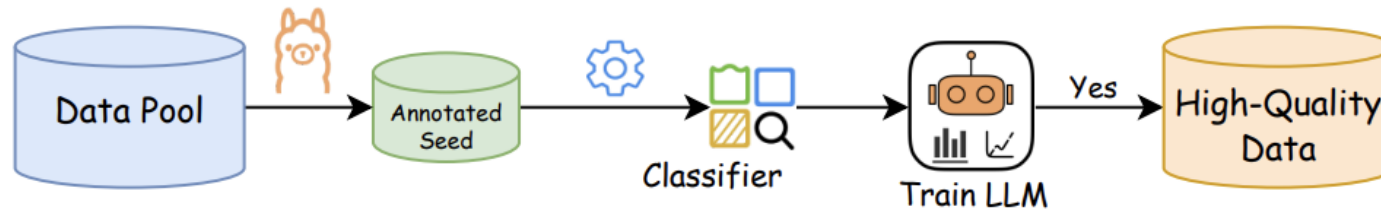
[IndicLLMSuite]

Usefulness of Data Cleaning



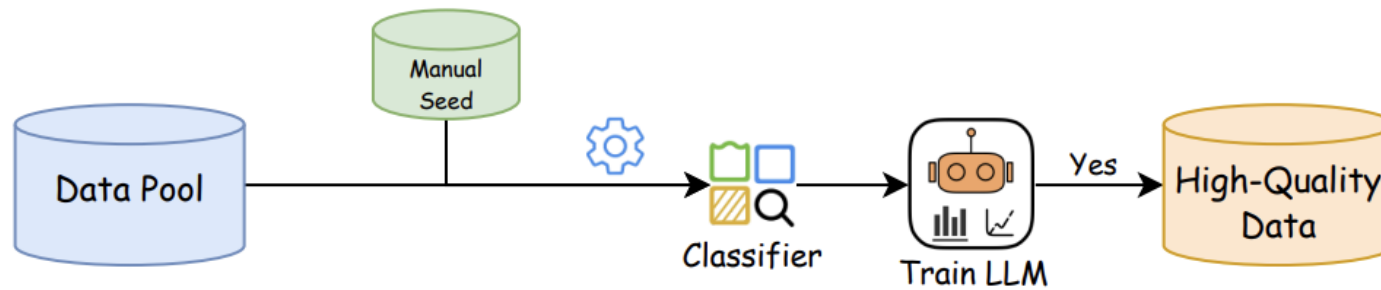
[FineWeb2]

Model Based Filtering



(a) LLM annotation-based high-quality filtering pipeline

*Get a better proxy
for quality than
simple heuristics.*



(b) Manual seed-based high-quality filtering pipeline

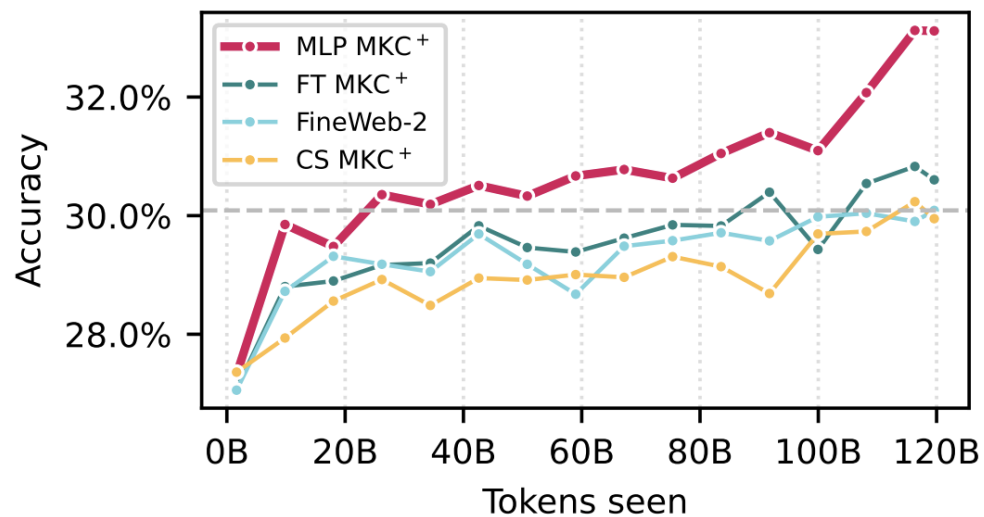
Can be SLLMs, or
simple n-gram
models or simple
fast-text based
classifiers

[UltraFineWeb]

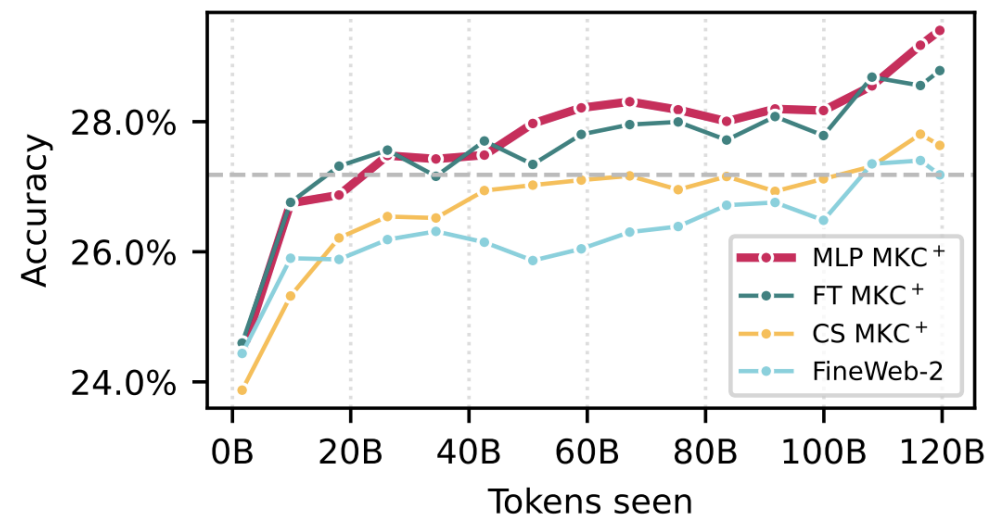
Model Based Filtering

Table 2: Comparison of inference costs for different model-based classifiers on 15T tokens

	LLM-based Classifier	fastText-based Classifier
GPU Used	✓	✗
CPU Used	✓	✓
Processing Time (Hours)	6,000	1,000



(b) Chinese (CMMLU)



(c) German (MMLU)

Model Based Filtering

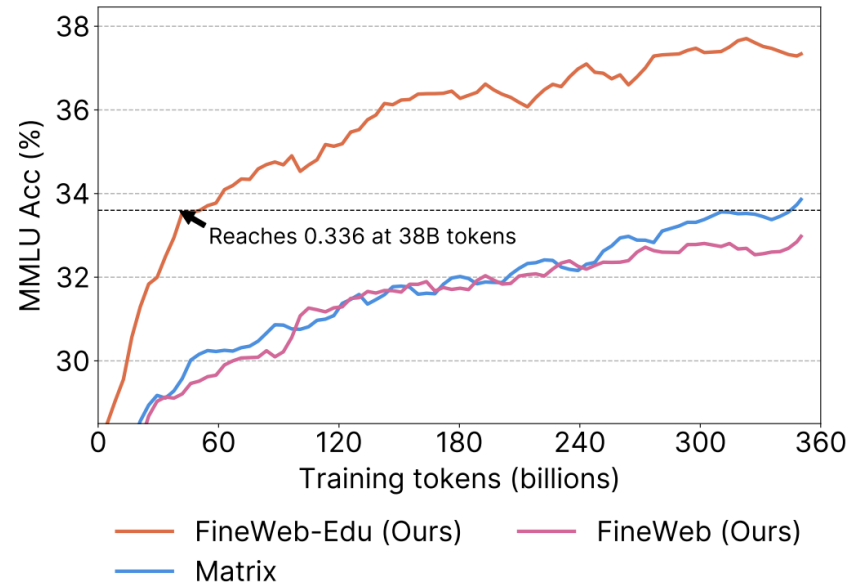


Figure 11: **Performance Comparison on MMLU.** FineWeb-Edu achieves a 33.6% accuracy on the MMLU benchmark at only 38 billion tokens, significantly outperforming Matrix (second best on the metric), which reaches similar accuracy at 300 billion tokens.

Model based filtering can help in filtering documents from targeted domains (like STEM, Reasoning, Legal, etc)

Significant downstream performance boosts compared to original FineWeb2

Great, so what after Data?

The Multilingual CPT Recipe

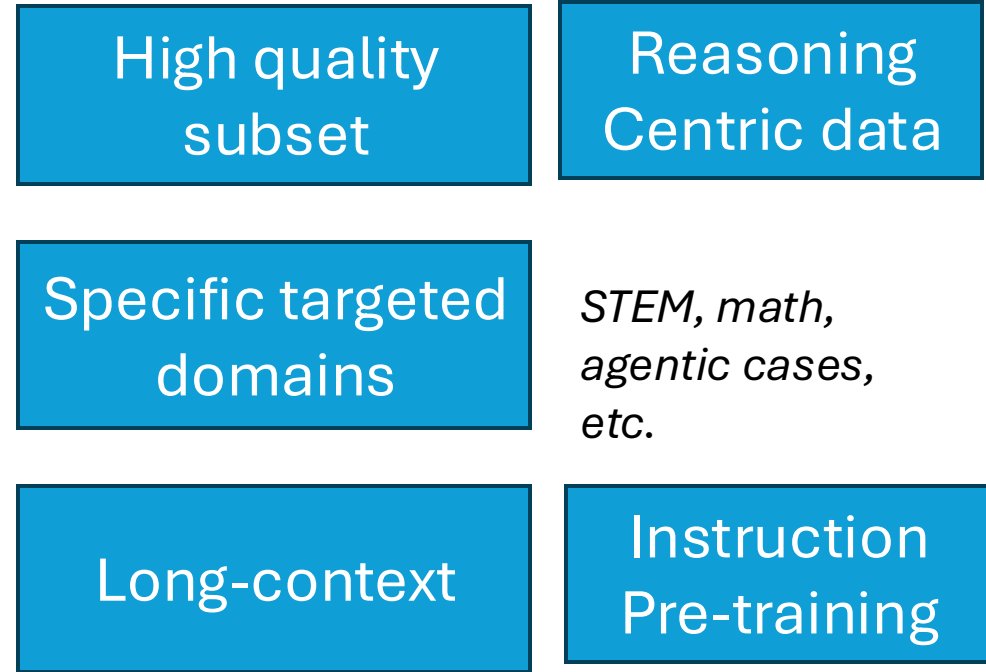
[Sailor2]



Initial CPT

Train on document-level data

Finetuning on long, coherent sequences helps model learn and correlate different pieces of knowledge



***Late-stage CPT Annealing/
Cooldown***

Balanced distribution across languages

Optimize learning rates and schedules for CPT

Does English performance get impacted by CPT?

Model	Perplexity	Linguistic Tasks	Knowledge Tasks
Base	2.40	50.36	79.28
CPT w/o English	3.68	29.38	44.23

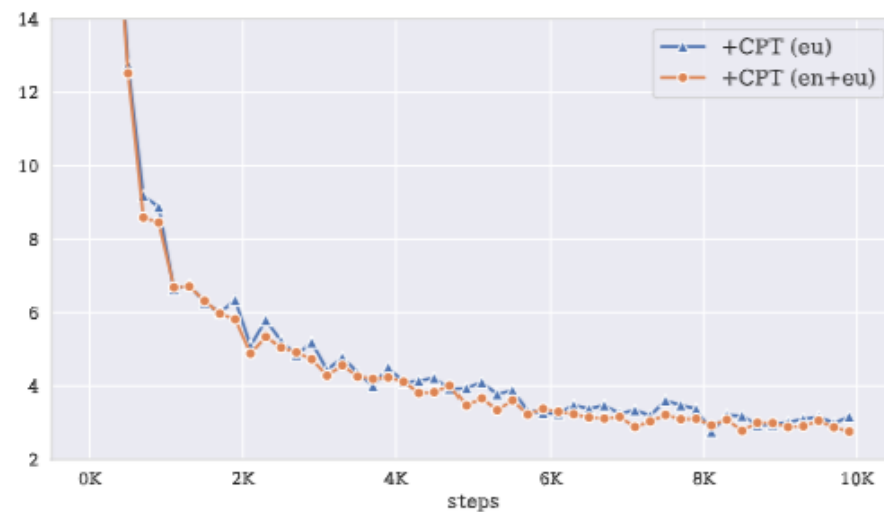
CPT on target language can cause catastrophic forgetting and degrade English performance

Is retaining English language capabilities and competence important?

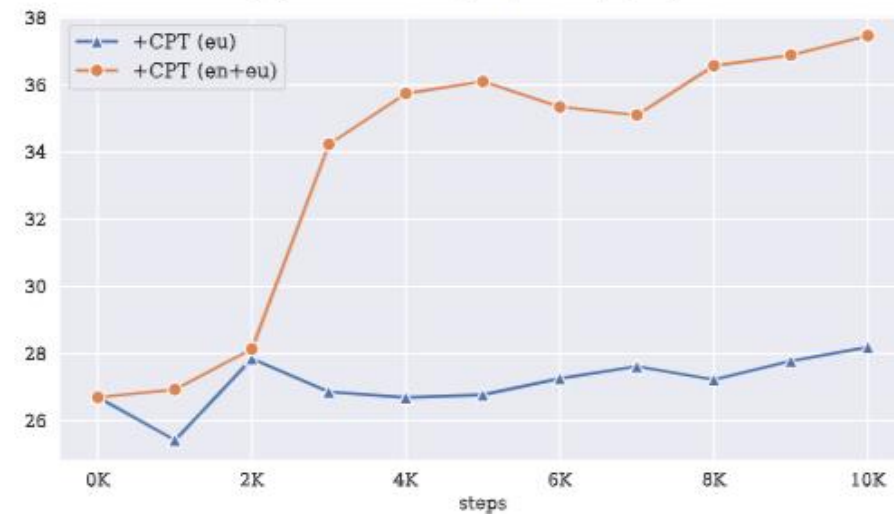
- We might want to use the same model for English and the new language
- Can drop in English capabilities impact other languages?

Is retaining English language capabilities and competence important? [EmergentCPT]

Drop in English capabilities can impact other languages also



(a) Validation perplexity (eu)



(b) Downstream accuracy (eu)

***CPT only on target language can improve language modeling,
not downstream tasks***

What happens if CPT does not include English? [EmergentCPT]

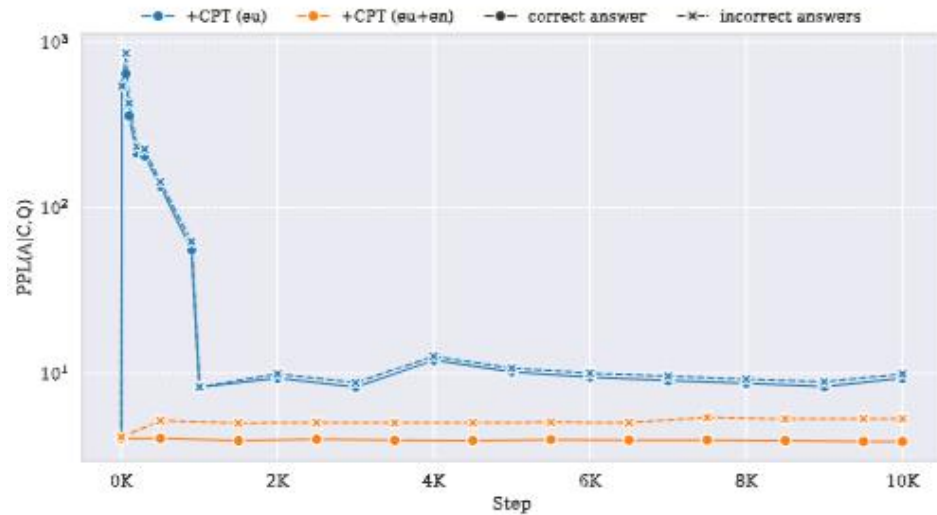


Figure 3: **Perplexity of choice labels on Basque downstream tasks for Llama 2 7B.** The variant without

Distribution shift from switching to a new language can result in a catastrophic forgetting of the ICL capabilities of the model

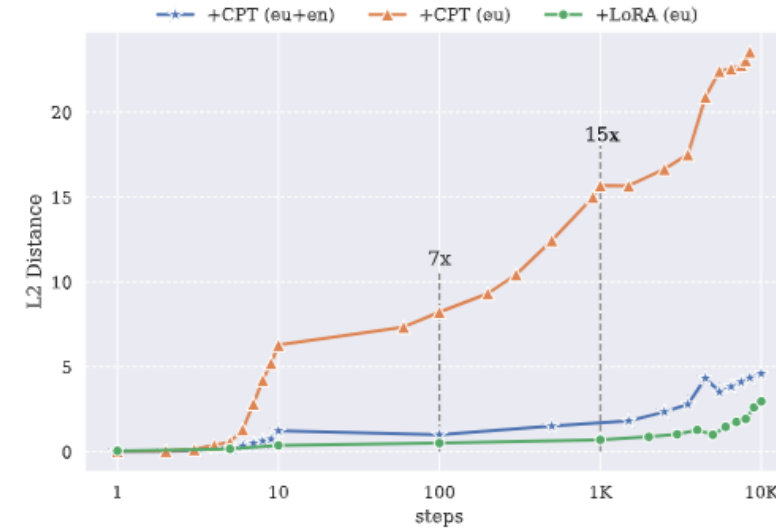


Figure 4: **Average layer-wise L2 distance of model parameters** from the initial Llama 2 7B model throughout full-parameter CPT and using LoRA. The CPT variant

Strong Parameter shift without English

Using LoRA for CPT [EmergentCPT]

LoRA and other PEFT methods can be used to reduce catastrophic forgetting

	PPL	Dwn.	Cop.
Llama-2 (7B)	23.64	27.43	44.67
+ CPT (eu+en)	3.35	34.14	43.43
+ CPT (eu)	3.58	28.89	20.12
+ LoRA (eu)	3.68	28.03	39.61

- Can help reduce perplexity
- Improved performance compared to no-English CPT
- Still lags CPT with English

PEFT methods lack sufficient capacity to make use of CPT data

Data Replay: Adding English to CPT Mix

- **Full Replay:** Include English in the CPT mix throughout → overhead
 - It can be sampled at less ratio than the target language (typically 20-30%)
- **Curriculum Replay:** Include English only during the initial part of CPT

	PPL	Dwn.	Cop.
Basque (eu)			
Llama 2 (7B)	23.64	27.43	44.67
+ CPT (full)	3.35	34.14	43.43
+ CPT (curr)	3.08	35.12	42.94
Llama 2 (13B)			
Llama 2 (13B)	13.66	29.52	49.23
+ CPT (full)	2.82	42.52	47.80
+ CPT (curr)	2.65	42.42	46.33
Arabic (ar)			
Llama 2 (7B)	4.36	32.45	44.67
+ CPT (full)	2.09	34.34	32.60
+ CPT (curr)	2.00	34.53	39.66
Indonesian (id)			
Llama 2 (7B)	6.27	26.65	44.67
+ CPT (full)	3.25	30.79	30.79
+ CPT (curr)	3.14	29.09	31.03

Data Replay helps improve performance in all aspects

Curriculum based CPT can be as effective as full CPT

+ Achieves significant savings in compute

Data Replay also helps regain English performance

Model Merging Original PT and CPT models

- Popular way to improve model performance and making it robust [EMA]
- Shows to work well for SFT and domain adaptation [SMOLLM3,SWAD]
- Model Merging for CPT
 - Merge CPT with original checkpoint [Tower+]
 - Exponential Moving Average (EMA) of checkpoints [EmergentCPT]
- **No need for English CPT**

$$\theta_t = \begin{cases} \theta_t^I & \text{if } t \leq 0 \vee t \bmod \eta \neq 0 \\ \alpha \theta_{t-\eta} + (1 - \alpha) \theta_t^I & \text{otherwise} \end{cases}$$

(assign higher weight to older checkpoint)

	PPL	Dwn.	Cop
Basque (eu)			
Llama 2 (7B)	23.64	27.43	44.67
+ CPT (eu+en)	3.35	34.14	43.43
+ CPT w/ EMA (eu)	2.98	34.89	42.66
Llama 2 (13B)			
Llama 2 (13B)	13.66	29.52	49.23
+ CPT (eu+en)	2.82	42.52	47.80
+ CPT w/ EMA (eu)	2.71	41.39	42.99
Arabic (ar)			
Llama 2 (7B)	4.36	32.45	44.67
+ CPT (ar+en)	2.09	34.34	32.60
+ CPT w/ EMA (ar)	2.03	33.36	42.76
Indonesian (id)			
Llama 2 (7B)	6.27	26.65	44.67
+ CPT (id+en)	3.25	30.79	30.79
+ CPT w/ EMA (id)	2.97	29.11	33.34

EMA is very competitive with full/curr CPT

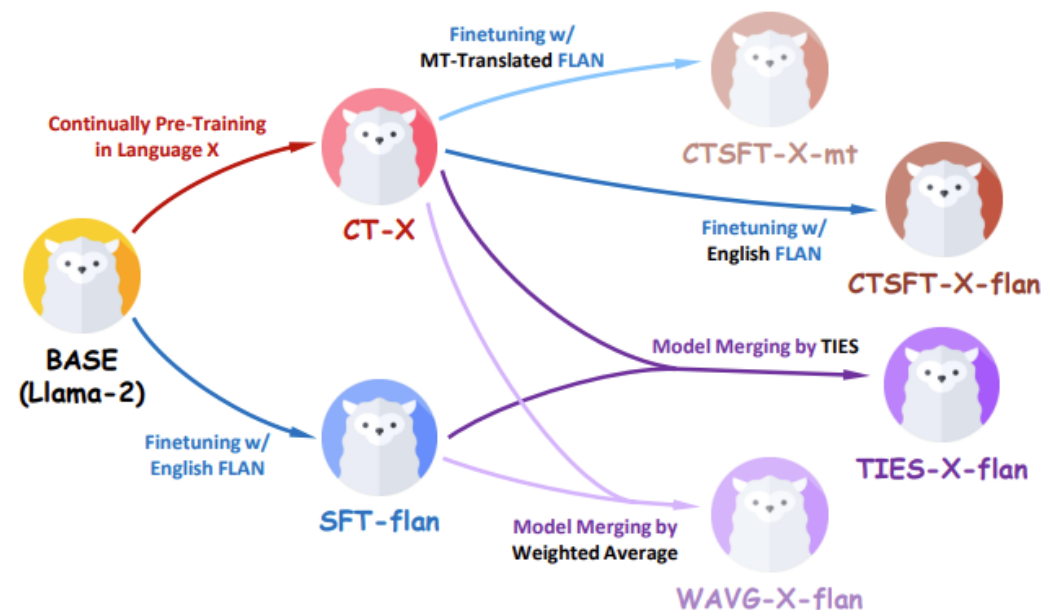
[EmergentCPT]

Model Merging CPT and SFT models

Model merging can also be used to combine language and task skills

Merge CPT and English SFT models

- Model Merging can outperform CPT-then-SFT in most cases
- CT-then-SFT can be preferred in resource constrained scenarios



	Task	Lang.	Tamil	Telugu	Odia	Bengali	Tibetan	Uyghur	Mongolian	Average
BASE			28.15	18.83	26.64	25.47	13.49	13.34	11.57	19.64
SFT-flan	✓		29.19	17.28	25.21	24.84	23.29	22.27	19.86	23.13
CT-X		✓	52.18	34.67	47.93	30.77	13.52	14.80	11.09	29.28
CTSFT-X-mt	✓	✓	50.57	32.90	30.14	38.40	33.85	24.85	19.57	32.90
CTSFT-X-flan	✓	✓	53.95	<u>37.96</u>	44.56	42.19	42.36	<u>49.46</u>	24.29	42.11
WAVG-X-flan	✓	✓	<u>57.56</u>	37.58	<u>53.59</u>	37.19	<u>44.30</u>	42.64	<u>31.09</u>	<u>43.42</u>
TIES-X-flan	✓	✓	58.46	39.50	56.49	<u>40.31</u>	47.86	52.43	32.56	46.80

[ModelMergingLRL]

Model Merging is a useful tool in the language adaptation toolkit for efficient adaptation

Expansion to large number of languages

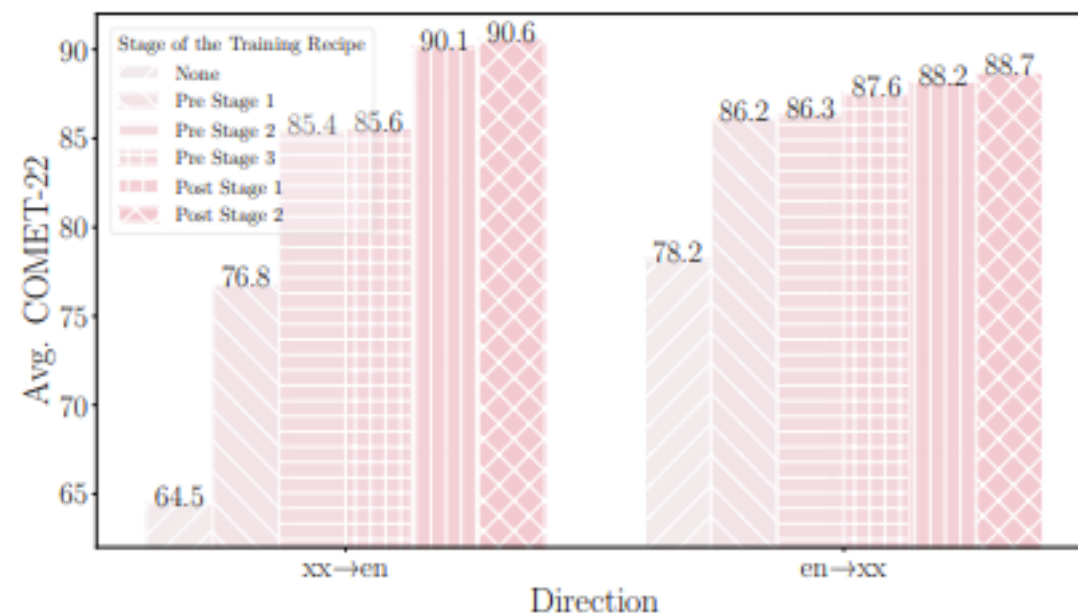
3 phase pre-training combining Full and LoRA CPT

1. Full CPT for all languages
2. Finetuning language-group specific LoRA modules
3. Finetune the language-group specific LoRA modules on parallel data

Language-specific modules really help handle a large number of languages

15% more parameters due to LoRA → increased capacity helps

Able to support CPT of 50 languages



Improving Cross-lingual Transfer in Pre-training

Why?

- Help improve cross-lingual alignment
- Make knowledge available in English in the target languages
- Help translation task
- Using Parallel/Translated Data
- Using Romanized Representation

Using Parallel Data

[Tower, Palm2, PolyLM, OpenHathi, ALMA]

Using human-written parallel data

- Train on document/paragraph pairs → very little availability
- Train on sentence pairs → modest availability depending on language pair

<lang1> [text] <lang2> [text]
<lang2> [text] <lang1> [text]

Training Schedule

- Monolingual PT followed by Parallel PT (for translation LLMs) *[ALMA]*
- Joint Monolingual and Parallel data (upsampled) PT *[Tower, Palm2]*
- Parallel PT followed by Monolingual PT *[OpenHathi]*

Parallel data is useful for translation task

[Tower, ALMA, OpenHathi, InciBiling]

High-quality and limited parallel data is sufficient

Use Mono.	Parallel Data Quality	Avg. xx→en		Avg. en→xx	
		BLEU	COMET	BLEU	COMET
✗	✗	26.43	77.16	13.86	68.88
✗	Random	28.24	78.69	19.68	73.89
✗	Filtered	28.39	78.94	19.56	74.35
✗	HW	29.39	80.00	22.17	76.52
✓	✗	28.49	80.32	26.35	84.73
✓	Random	32.47	83.02	26.98	83.15
✓	Filtered	32.32	83.03	27.38	83.98
✓	HW	34.02	84.08	29.89	86.49

Believed to improve LLM ability to understand and generate multilingual text [Palm2]

No systematic results on utility of parallel data in pre-training on downstream tasks

Previous work on encoder and encoder-decoder PLMs *[PrimerPMLM]*

- Improves retrieval and translation quality significantly
- Modest improvement for cross-lingual transfer in different tasks

Machine Translated Data for Monolingual Pre-training

[IndicMonoDoc, IndicLLMSuite, TransLLMWeb, MTPretrainLessons]

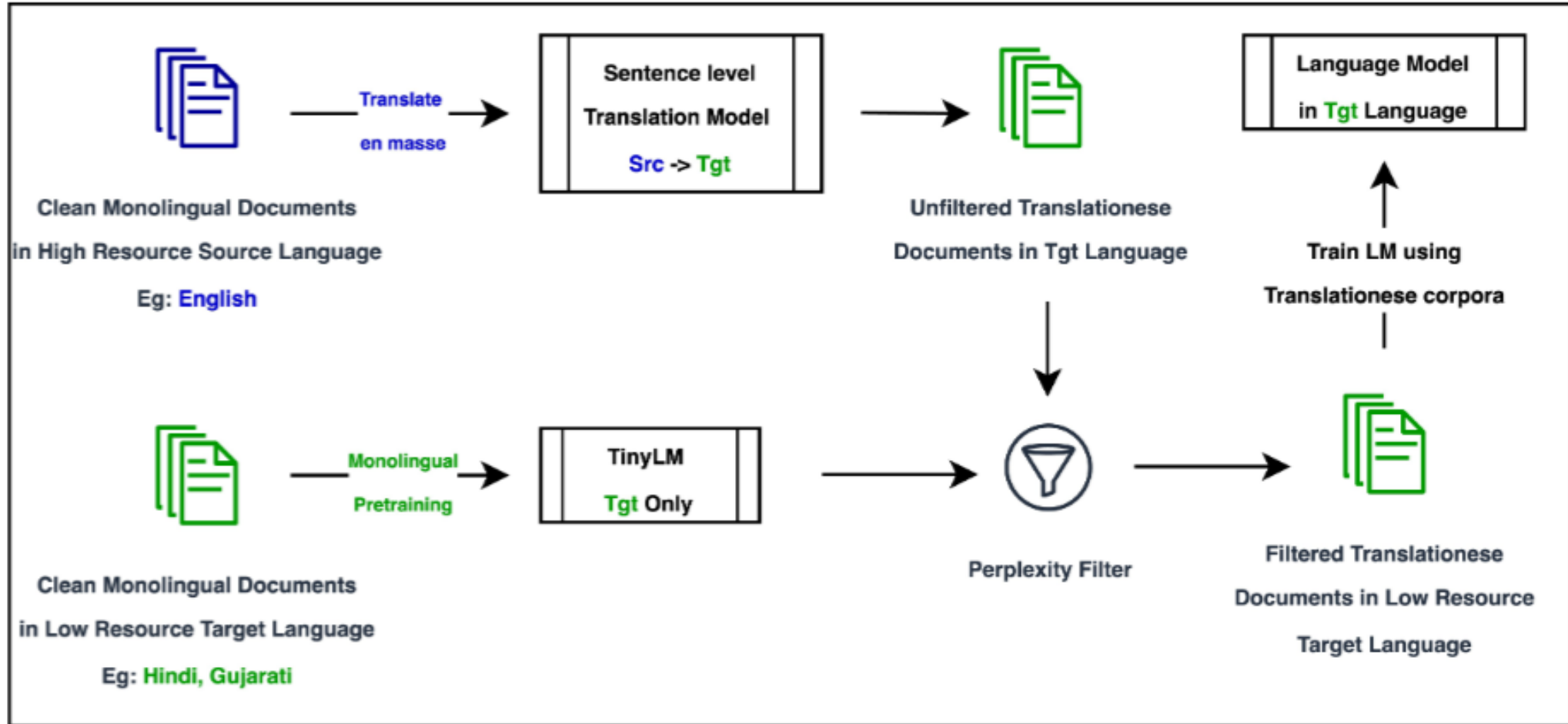
Why:

- Most languages do not have sufficient monolingual data to capture knowledge
- Transfer of knowledge in multilingual LLMs is challenging

Simple, brute-force solution: Translate English content to other languages

- Needs a decent MT model
- NMT as well as LLM-based translation perform similarly on downstream eval
- Can be most useful **high-quality, knowledge intensive corpora** like Wikipedia, FineWeb-EDU

Machine Translated Data for Monolingual Pre-training



[IndicMonoDoc]

Effect of MT data on Pre-training

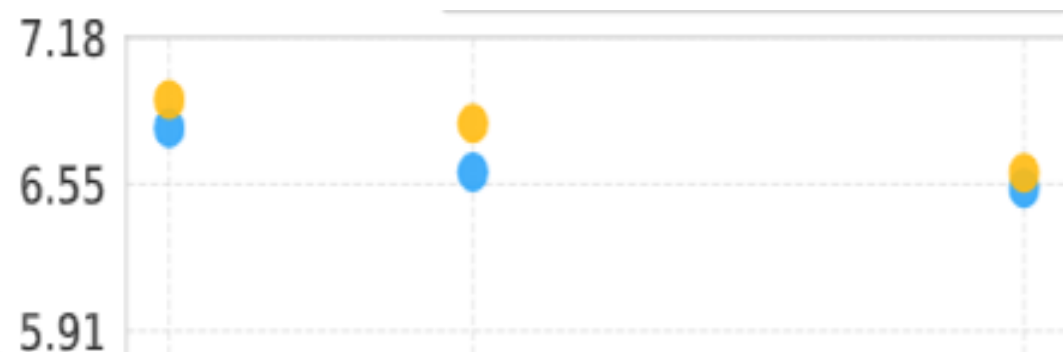
- Pre-training on **MT data competitive to native language documents** on many NLP tasks
- **Translation quality filtering + CPT on small native data improves quality**
- For small LMs, synthetic data might outperform original data

Model	NLU						NLG				
	iXNLI	bbc-a	iitp-mr	iitp-pr	midas	Avg.	Headline Gen.	Sentence Summ.	Question Gen.	Wikibio	Avg.
HI-clean	73.61	81.75	72.58	79.73	80.34	77.60	27.54	23.64	24.84	52.16	32.04
syn-HI_en-unfiltered	72.87	77.92	64.36	76.22	79.91	74.26	27.29	22.93	24.22	50.14	31.14
syn-HI_en-unfiltered+10%	74.63	78.36	67.75	77.46	80.17	75.67	26.98	23.20	24.76	51.34	31.57
syn-HI_en-filtered	74.75	81.06	69.03	78.58	79.73	76.63	27.15	23.10	24.41	49.88	31.13
syn-HI_en-filtered+10%	74.49	80.94	71.61	79.92	80.64	77.52	27.87	24.23	24.87	51.18	32.04

[IndicMonoDoc]

Model size Scaling is effective even for MT data

MT data is less effective for medium/low-resource languages with lower MT quality



[MTPretrainLessons]

Limitations to using MT data

- Models trained on MT data weak on linguistic proficiency and cultural knowledge
- Additional CPT on small native data improves quality can help overcome these deficiencies

	Native LLM	MT LLM	MT+ Native CPT LLM
Linguistic Proficiency	79.41	65.13	74.79
Cultural Awareness	62.62	48.84	57.61

[TransLLMWeb]

Linguistic competency test for French on fr-grammar/vocab tests

Cultural awareness tests for Indonesian on COPAL-ID testset

MT data can be successfully used for CPT of an English-heavy LLM

Data	Marathi		Gujarati	
	Gemma 2B	Llama-3 8B	Gemma 2B	Llama-3 8B
Base model	178.898	66.740	71.136	2.839
clean	37.599	11.196	10.350	2.312
synthetic-unfiltered	92.813	15.697	10.941	2.816
synthetic-filtered	104.148	14.622	10.150	2.236

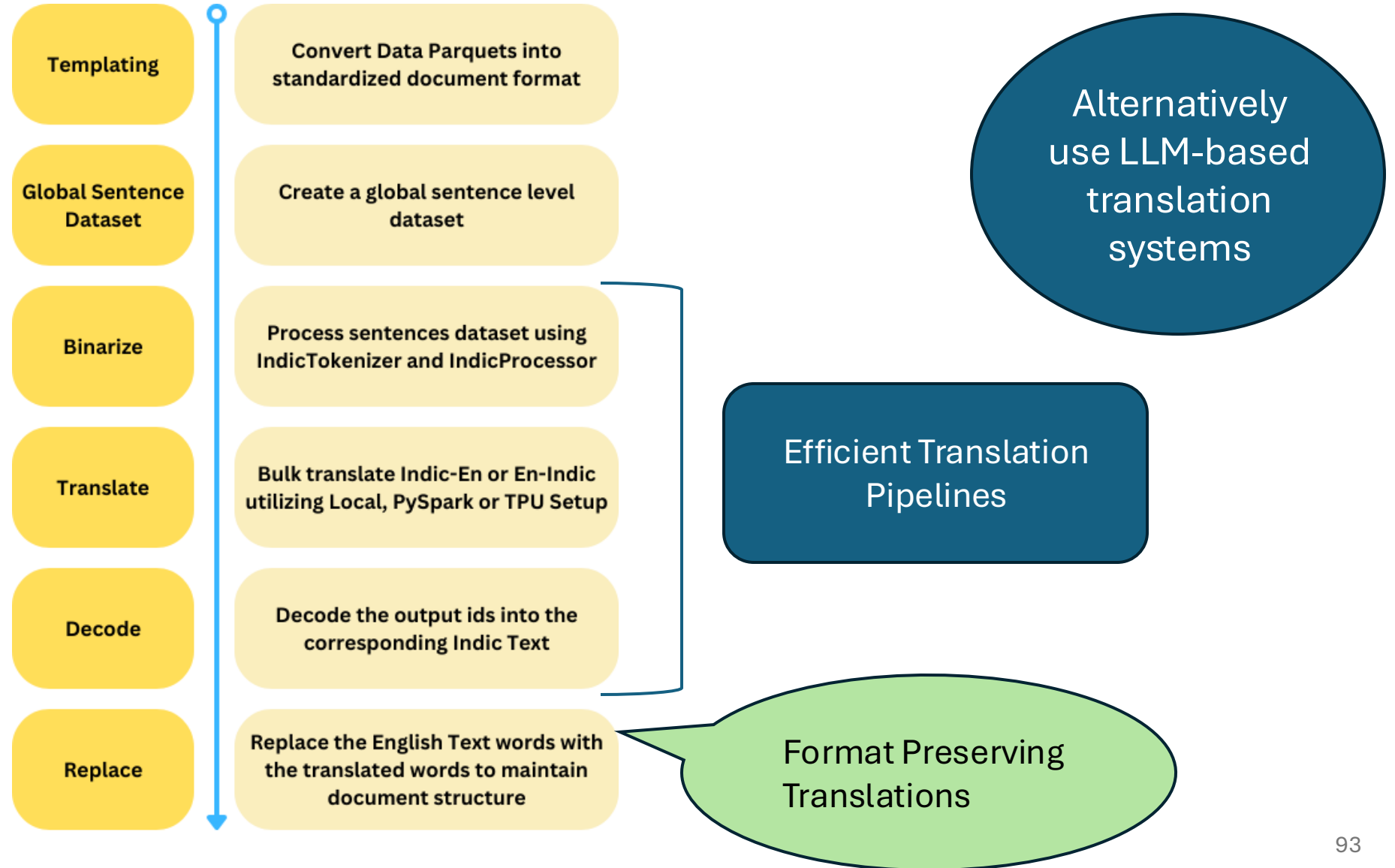
LM Perplexity

[IndicMonoDoc]

Data	Marathi		Gujarati	
	Gemma 2B	Llama-3 8B	Gemma 2B	Llama-3 8B
Base model	90.89 ± 0.005	95 ± 0.009	83.84 ± 0.002	92.69 ± 0.018
clean	90 ± 0.014	97.17 ± 0.013	87.79 ± 0.012	93.33 ± 0.0132
synthetic unfiltered	89.66 ± 0.012	95.38 ± 0.016	83.88 ± 0.009	92.81 ± 0.01
synthetic filtered	86.67 ± 0.016	96.15 ± 0.011	84.10 ± 0.013	92.94 ± 0.007

Accuracy on few-shot sentiment analysis

Large Scale Pipelines for Translation



Pre-training on Romanized Representation

[RomanSetu, IndicLLMSuite, TransliCo]

Non-Roman script challenges: High-fertility/data loss, expensive vocab extension

<s> चारोॆं अॆंतरिक्ष याॆतॆरी बेॆंगलुॆरु मेॆं भाॆरतीॆय अॆंतरिक्ष अनुॆसॆंधाॆन संग<0xE0><0xA4><0xA0>न (<0xE0><0xA4>
<0x87>सरो) की अॆंतरिक्ष याॆतॆरी सुॆविॆधा मेॆं पॆरशॆक्षण लेॆ रहे हैॆं।<0x0A>

(130 tokens)

<s> chaaron antariksh yaatree bengaluru mein bhaarateey antariksh anusandhaan sangathan (isaro) kee antariksh ya
atree suvidha mein prashikshan le rahe hain.<0x0A>

(63 tokens)

Why use romanized data

- Many languages are additionally written in romanized script
- The same script as English → better alignment?

Transiteration choices

- Natural transliteration
- Fixed Romanization schemes

Language	N	R
Gujarati	18.44	3.39
Hindi	7.36	2.98
Malayalam	12.85	5.04
Marathi	8.91	3.64
Tamil	12.11	4.89

Romanized fertility more than 2x lower than native script fertility

[RomanSetu]

Language	E - N	E - R
Gujarati	0.39	0.47
Hindi	0.40	0.50
Malayalam	0.40	0.46
Marathi	0.44	0.48
Tamil	0.44	0.43

Romanized representations are better aligned to English than native script representations

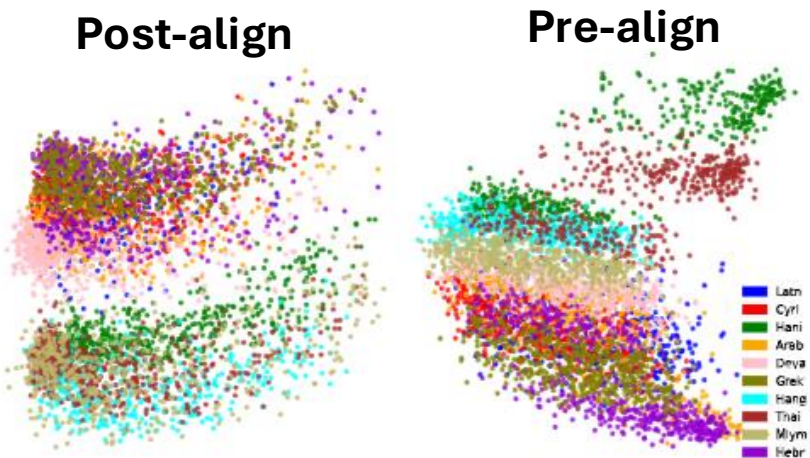
Aligning the romanized representations explicitly

[TransliCo]

Fine-tune an LLM by contrasting sentences in its training data and their Roman transliterations

Aligns the representations across scripts

Improves cross-lingual transfer on a variety of tasks



	Taxi1500			NER			POS		
	XLM-R	Glott500	FURINA	XLM-R	Glott500	FURINA	XLM-R	Glott500	FURINA
Latn	22.5	<u>52.6</u>	59.8	60.3	<u>66.1</u>	67.3	68.1	<u>74.4</u>	75.7
Cyrl	30.2	<u>59.8</u>	63.6	51.8	<u>65.3</u>	66.2	66.7	<u>79.3</u>	79.5
Hani	66.6	<u>68.2</u>	70.1	23.1	<u>22.2</u>	21.9	<u>22.2</u>	35.5	18.2
Arab	48.5	<u>60.8</u>	66.5	45.0	<u>53.4</u>	57.7	65.8	<u>68.8</u>	69.3
Deva	49.5	<u>66.6</u>	73.2	56.9	<u>56.2</u>	58.9	58.3	<u>59.8</u>	60.8
Other	49.5	<u>59.5</u>	65.2	45.2	<u>50.4</u>	50.4	65.9	<u>68.8</u>	67.1
All	26.7	<u>54.3</u>	61.0	55.3	<u>61.6</u>	62.8	65.6	<u>71.8</u>	71.9

Summary and Recommendations

- CPT helps improve **language competence and model’s “accessible” knowledge.**
- **High-quality, diverse and culturally relevant data** should be curated.
- **Critical to maintain English task performance** during CPT
 - Avoiding catastrophic forgetting: Data Replay, Adapters, Model Merging
- **Does parallel data improve cross-lingual transfer?**
 - Improves translation quality
 - Improving cross-lingual transfer, use of translated data requires further research
- Data augmentation methods **like romanization, code-switching** are helpful

Tutorial Outline

- Introduction
 - The Multilingual LLM Challenge
- Extending LLMs to new languages
 - Vocabulary Expansion
 - Continued Pre-training
 - **Post-Training**
- Multilingual Evaluation
- Summary & Future Directions

Post Training

Language Fluency
(Text Completion)

Pre-trained Model



Post Training



Post-trained Model

Supervised Instruction Tuning

Preference Alignment

Reasoning Capabilities

...

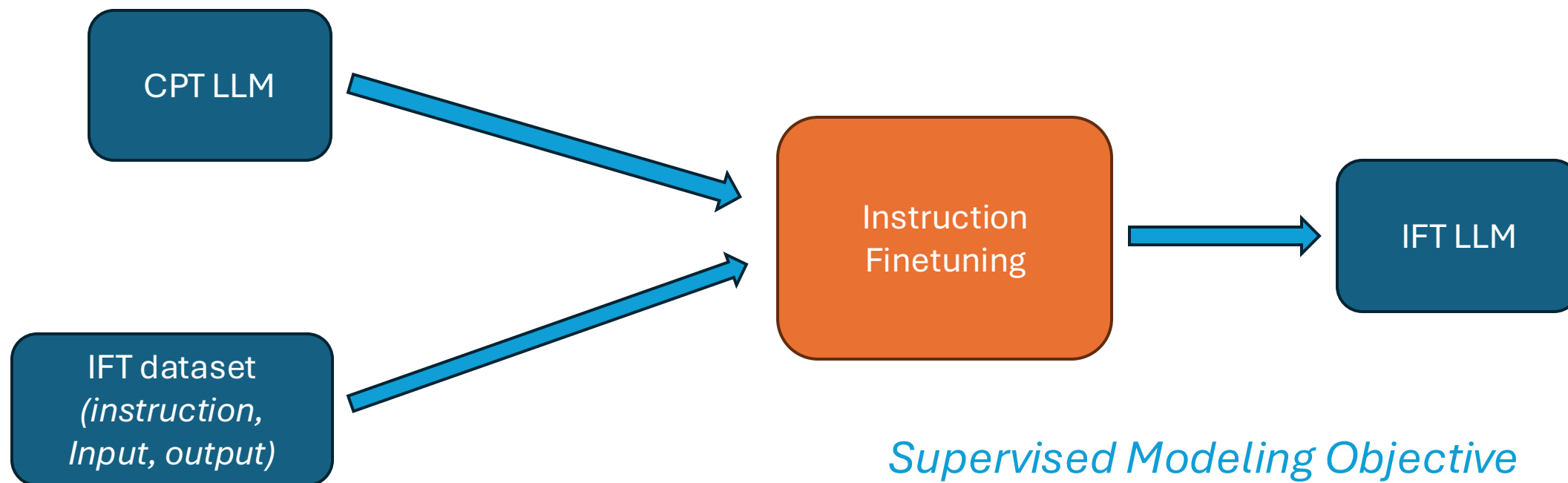


Chat Capabilities

Outline

- Supervised Instruction Tuning
 - Data Creation Strategies
 - Training Strategies
 - Ablations
- Reinforcement Learning
 - Intro to RL
 - RL for preference alignment
 - RL for Reasoning

Supervised Instruction Tuning



Train on in-language IFT dataset

Sources of IFT dataset

Quality and diversity of IFT dataset

Supervised Modeling Objective

$$\ell_{\text{CE}}(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{j=1}^{|\mathcal{V}|} y_j \log(\hat{y}_j)$$

Only computed over the output tokens

$$\mathcal{L}_{\text{SFT}} = \frac{1}{N} \sum_{i=1}^N \ell_{\text{CE}}(\mathbf{y}_i, \mathcal{M}_{\theta}(\mathbf{x}_i))$$

To retain English task performance

Include English in the IFT training

Instruction Tuning Tasks

Variety of tasks/objectives to improve non-English performance

Generating IFT Data

Auxiliary Tasks

Transforming IFT
Datasets

- English Data IFT
- In-language IFT with Machine Translated Data
- Locally/Culturally relevant IFT data
- Synthetic Data Generation
- Parallel Data
- Monolingual Data
- Romanized IFT Data
- Cross-lingual Thought Data
- Cross-lingual IFT Data
- Code-switched IFT Data

Let's look at these tasks in detail

Generating IFT Data

Using English IFT Dataset

- Instruction tune the model on English instruction dataset
- Evaluate on non-English data → Zero-shot cross-lingual evaluation
- Instruction tuning on English important to retain English capabilities
- Important for performance in other languages also

Representative English IFT datasets

- Tulu3 Mixture
- UltraChat
- FLAN
- ORCA

Zero-shot Results

[MultilingualPostTrainer]

Summarization (Rouge-L)

Aya 7B		Aya 35B	
ar	0.24	ar	0.26
en	0.29	en	0.33
es	0.20	es	0.23
fr	0.25	fr	0.29
ja	0.31	ja	0.38
ko	0.14	ko	0.20
pt	0.26	pt	0.27
zh	0.26	zh	0.26
Avg.	0.25	Avg.	0.28

Reference English performance

Math (Accuracy)

Aya 7B		Aya 35B	
ar	0.00	ar	0.60
en	0.49	en	0.76
es	0.36	es	0.27
fr	0.19	fr	0.63
ja	0.18	ja	0.61
ko	0.26	ko	0.54
pt	0.12	pt	0.68
zh	0.23	zh	0.63
Avg.	0.23	Avg.	0.59

Emergent zero-shot cross-lingual transfer as we increase model parameters

Using Machine Translated IFT Dataset

- Translate English instruction tuning datasets into the language
- Fine-tune model on translated dataset

Task	BeleBele QA	MKQA	XL-Sum
	<i>Accuracy</i>	<i>F1</i>	<i>Rouge-L</i>
English IFT	45.58	36.48	8.42
Language IFT	48.28	37.95	15.87

*Average performance across many languages; **[SDRRL]***

Instruction tuning on translated data outperforms English instruction-tuning

Creating Translated IFT Data

Choice of Translation Engine

- Off-the-shelf NMT systems [Airavat]: higher quality, particularly for low-resource
- GPT [Okapi]: can do translation taking the entire context of input/output
- Hybrid Approach [LlmByndEng]: Do one of the above depending on language's translation quality

	#langs.	avg. chrF	avg. BLEU
ChatGPT (0-shot)	203	32.3	16.7
ChatGPT (5-shot)	203	33.1	17.3
GPT-4	20	44.6	24.6
NLLB	201	45.3	27.1
Google	115	52.2	34.6

Model	Human (General/Discourse)				
	News	Social	Fiction	Q&A	Ave.
Google	1.9/2.0	1.2/1.3	2.1/2.4	1.5/1.5	1.7/1.8
DeepL	2.2/2.2	1.3/1.1	2.4/2.6	1.6/1.5	1.9/1.9
Tencent	2.3/2.2	1.5/1.5	2.6/2.8	1.8/1.7	2.1/2.1
GPT-3.5	2.8/2.8	2.5/2.7	2.8/2.9	2.9/2.9	2.8/2.8
GPT-4	3.3/3.4	2.9/2.9	2.6/2.8	3.1/3.2	3.0/3.1

Comparison of various translation engines

Sentence-level

[ChatGptMT]

Comparison of various translation engines

Document-level

[ChatGptMT]

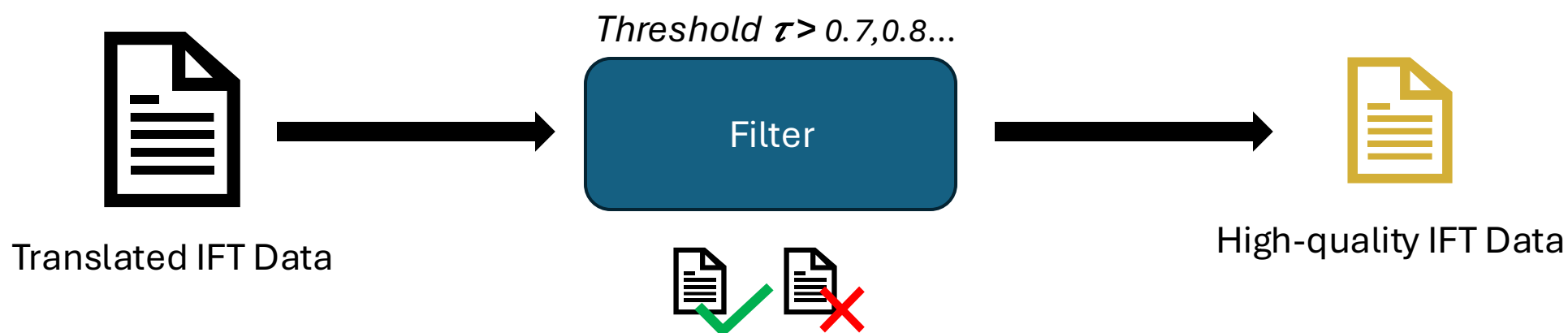
Creating Translated IFT Data (2)

Ensuring Translation Quality

Quality Filtering

High quality examples are important for instruction tuning

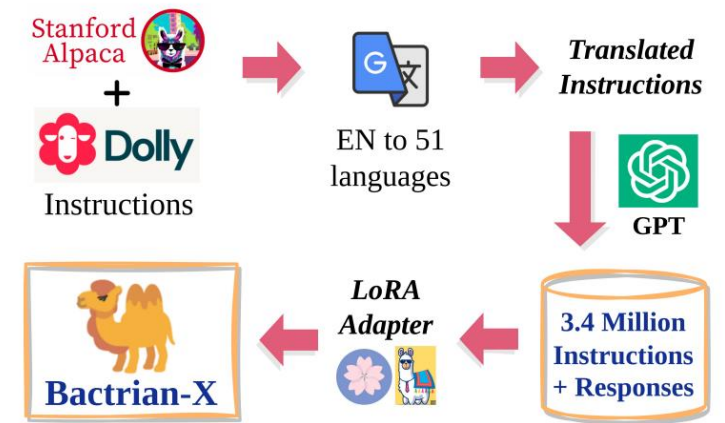
- Use an MT evaluation metric like COMET-QE to identify bad translations
- Rule-based filters to avoid code examples, etc. that are difficult to translate



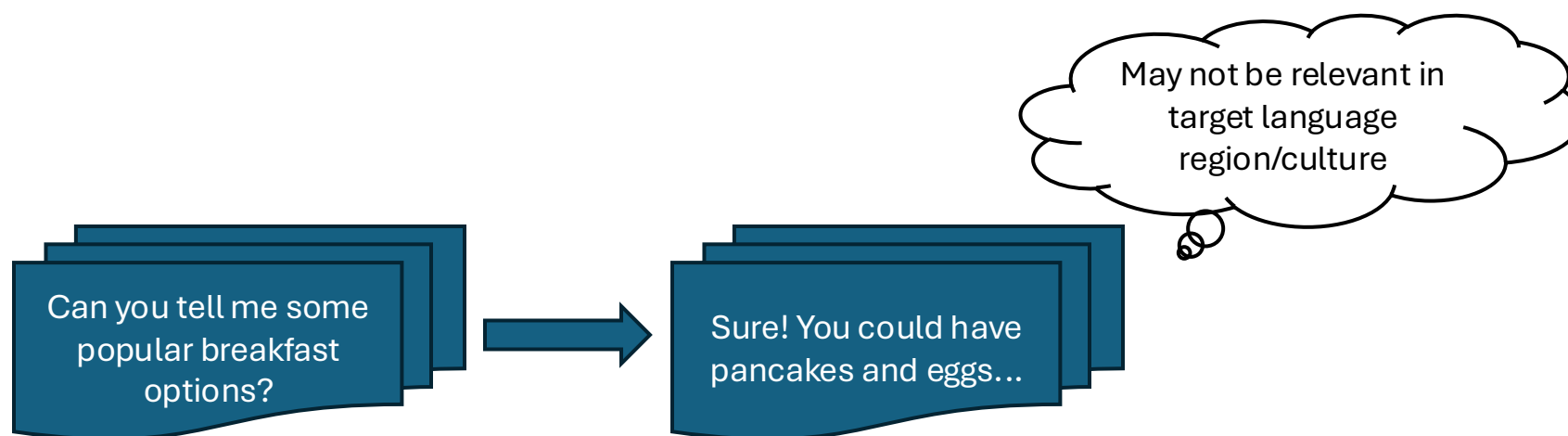
Creating Translated IFT Data (3)

What to Translate

- Instruction, Input, Output [Okapi, Airavat, xLLama, SDRRL]
- Input, Output [BLOOMZ]
 - English instruction is a common usecase
 - Models are good at English Instruction following
- Instruction, Input [BactrianX]
 - Give translated Instruction & Input
 - Generate response using GPT in the target language
 - Language/culture-specific examples



But we are just translating existing datasets – they might not be reflective of topics relevant to the language users

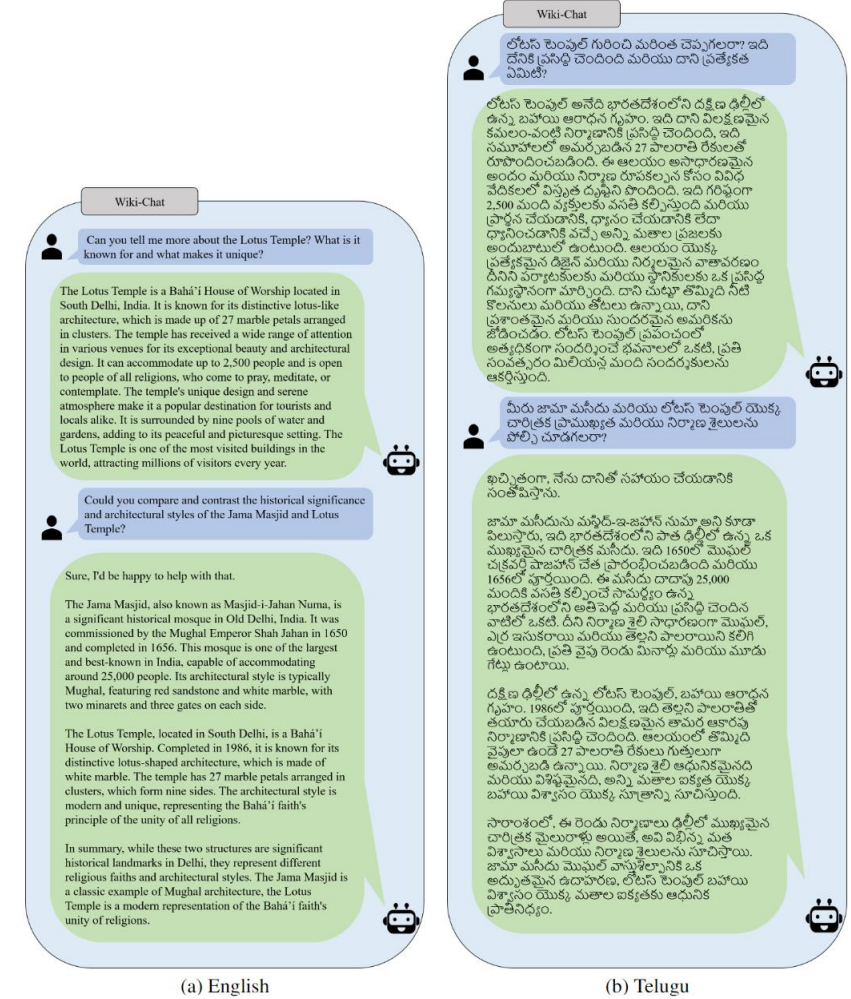
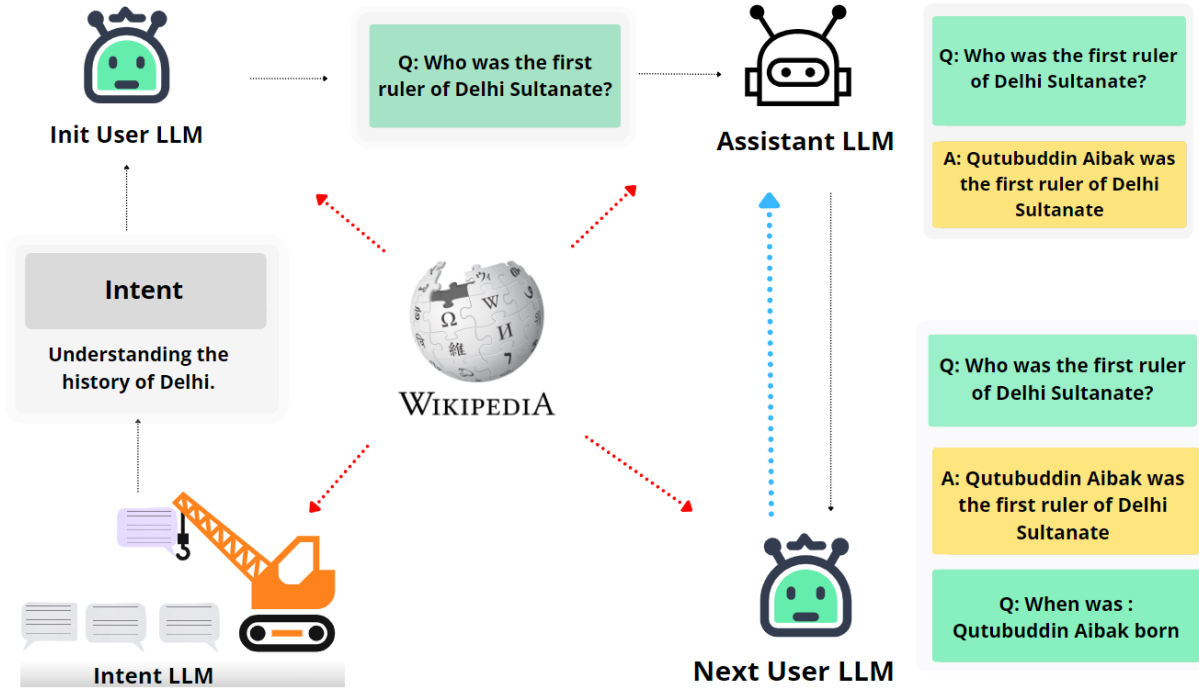


Can we synthetically generate SFT datasets?

Generating Culturally Relevant IFT Data through English

Creativity is the limit

Use multiple English LLMs along with Wikipedia context to simulate conversations on topics of interest



Translate the conversations from English to target language

Data generation in English?

→ LLMs are quite capable when it comes to English generation

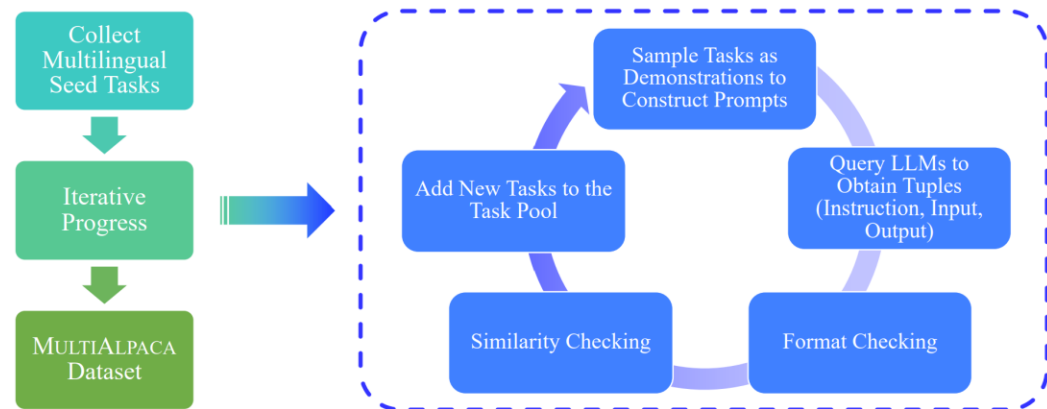
Can we generate data IFT data in the native language?

- *Limited language support in LLMs*
- *Quality of LLMs for multilingual generation*
- *Possibility of hallucinations, factual inaccuracies, cultural biases*

Curate small high-quality seed Instructions

Generates the entire examples from strong LLM like GPT in target language

Use methodologies like Instruction backtranslation



[PolyLM, SeaLLM, Updesh, AyaExpense]

Table 2: Generative task categories with synthesis methods, phases, and model configuration

Task Type	Synthesis Method	Phases	Qwen3-Mode
LOGICAL REASONING	Generate implicit inferences from text passages	(1) Direct inference generation	Reasoning
MULTI-HOP QA	Create questions requiring information synthesis across text segments	(1) Question generation (2) Answer generation	Reasoning
CREATIVE WRITING	Transform factual content into engaging narratives	(1) Generate creative piece (2) Generate eliciting prompt	Reasoning
MULTI-TURN DIALOGUE	Agentic workflows with 3-5 turn conversations between personas	(1) Generate dialog adhering to personas (2) Generate natural prompt	Non-reasoning
SUMMARIZATION	Generate summaries preserving key information across languages	(1) Direct summary generation	Non-reasoning
MACHINE TRANSLATION	Cross-lingual conversion maintaining cultural context	(1) Direct translation	Non-reasoning
CAUSAL REASONING	Identify and explain cause-effect relationships in text	(1) Direct causal analysis	Reasoning

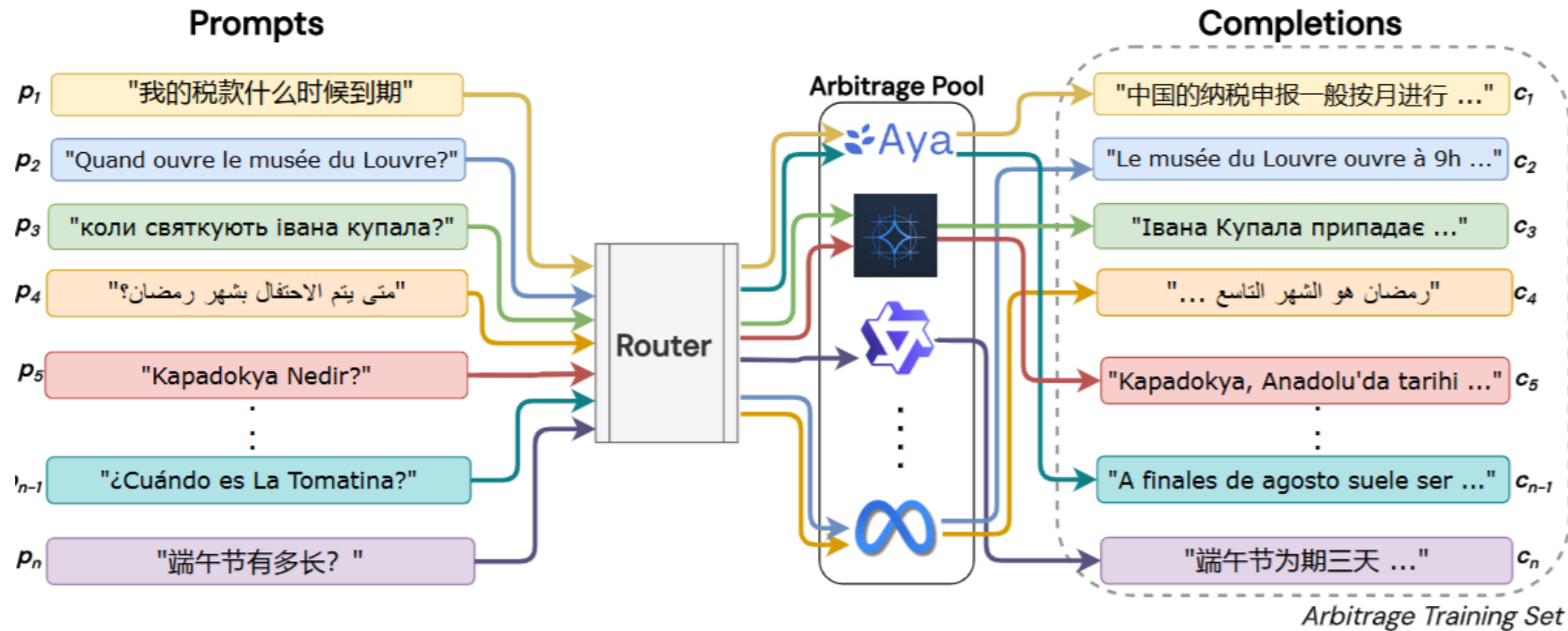
It is possible to synthesize diverse IFT data using strong multilingual LLMs along with high quality filtering with strong multilingual LLMs validated by human annotators

[Updesh]

Mix and Match depending on our use case

No single multilingual model is good for all scenarios.

Route the data generation to an appropriate LLM based on prompt – rule-based, learned router, etc.



Routing-based ensemble model can significantly improve final model performance

Auxiliary Tasks

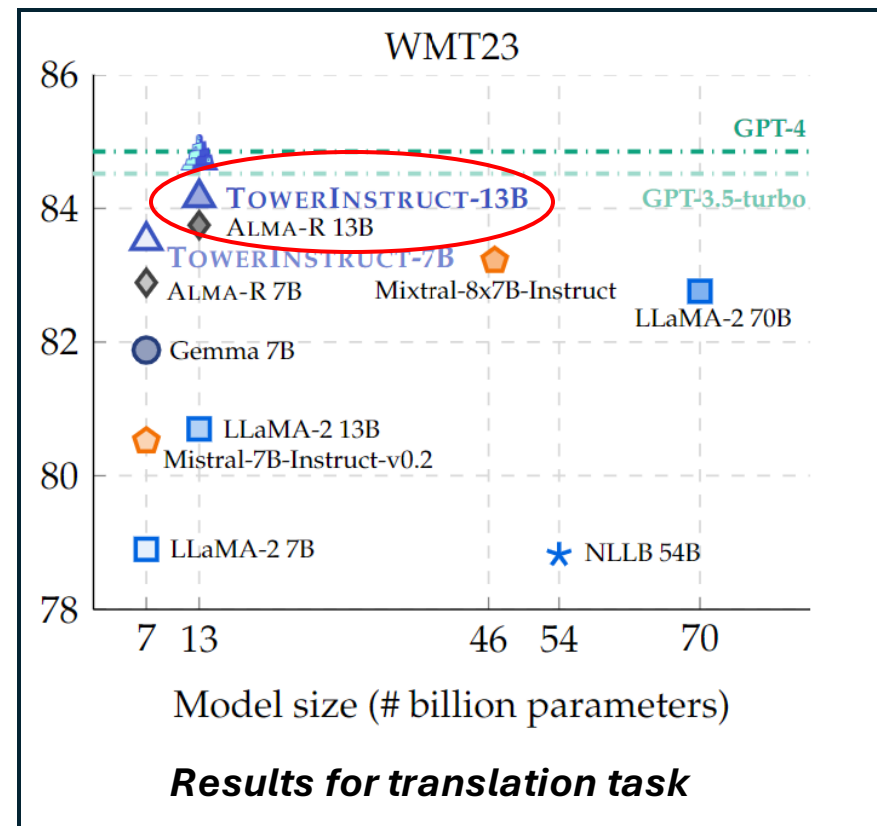
Can we align the language spaces for better cross-lingual transfer of knowledge and skills

Can parallel data or script uniformity help achieve the goal?

Using Parallel Data

Translation is a *special* task for multilingual IFT models

- Teaches the model to translate
- Aligns English and language representations better
- Improves performance on other downstream tasks
- Parallel data and translated IFT data both help



Instruction Data	XQUAD (exact match)	MLQA (exact match)	mLAMA (exact match)	XLSum (Rouge-1)
Alpaca-En	31.8	26.7	5.3	9.0
Alpaca-En+En-Zh	34.3	38.0	5.8	27.1
Alpaca-En+Alpaca-Zh	51.7	48.0	21.9	25.5
Alpaca-En+Alpaca-Zh+En-Zh	54.9	51.8	30.4	28.3

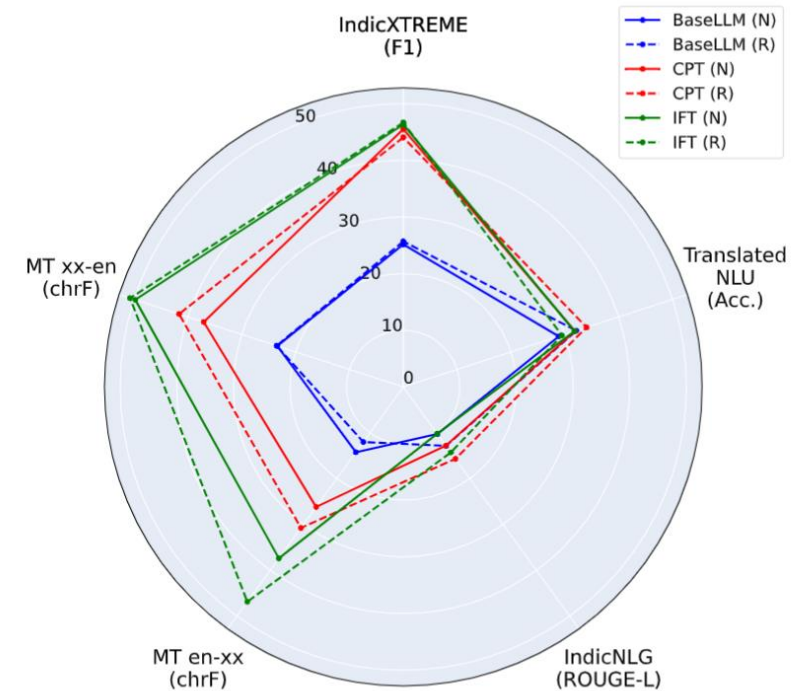
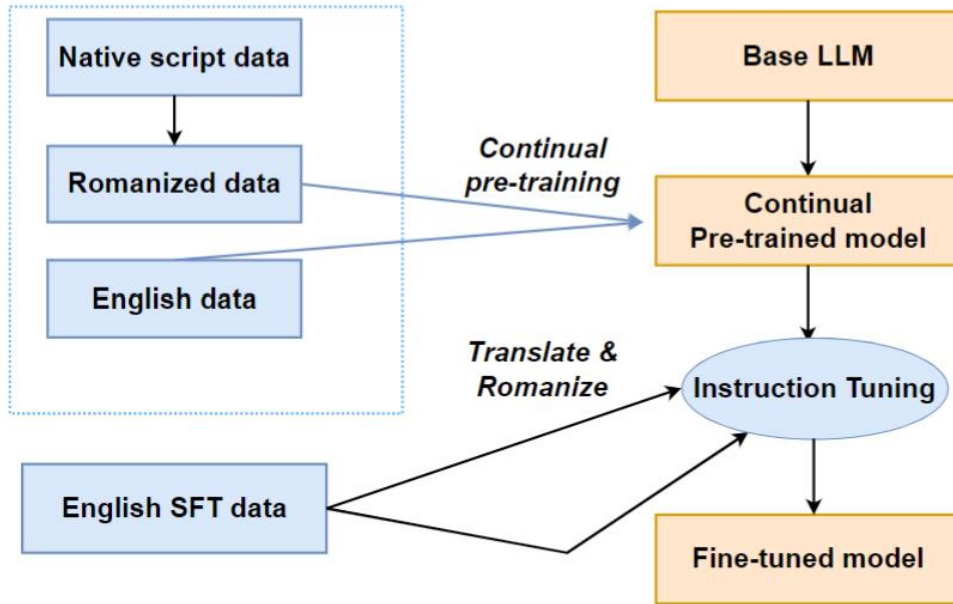
Results on Chinese for various Tasks

Instruction-tuning Data	Ar	Hi	Vi	Zh
Alpaca-En	16.1	13.7	34.1	26.7
Alpaca-En+En-Zh	33.6	35.1	42.2	38.0
Alpaca-En+Alpaca-Zh	33.1	35.1	50.1	48.0
Alpaca-En+Alpaca-Zh+En-Zh	37.0	42.3	50.8	51.8

Results for other languages on MLQA

Romanized Representation

Just like pre-training, use romanized representation for IFT too



- *NLG task performance improves with romanized data*
- *NLU task performance is on par, though more efficient*
- *Note: CPT with romanized data is crucial for Romanized Data IFT*

Can we further improve the fluency of the target non-English language text?

Can monolingual target language data help achieve the same?

Using Monolingual Data

- “Translationese IFT Data” → output language might not be fluent and high-quality
- Expose model to monolingual target language data during IFT
- Incorporate a task that helps model generate fluent output in target language

Task 1: Standard next-word prediction (CLM)

Switch between IFT and CLM objective in mini-batches

Task 2: Sentence Completion Task

Only IFT objective required

Question: Complete the following sentence in *Indonesian* according to its context.

Sang rubah coklat cepat

Answer: *Sang rubah coklat cepat melompati anjing malas.*

Transforming IFT Datasets

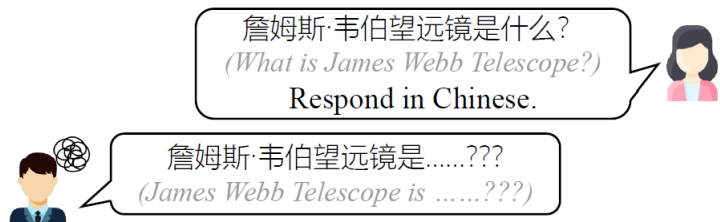
Models are still better in English

What if we ask in target language and use English as a proxy or vice versa?

Cross-Lingual Thought Prompting (XLT)

Ask the model to implicitly think in a different language

Monolingual QA

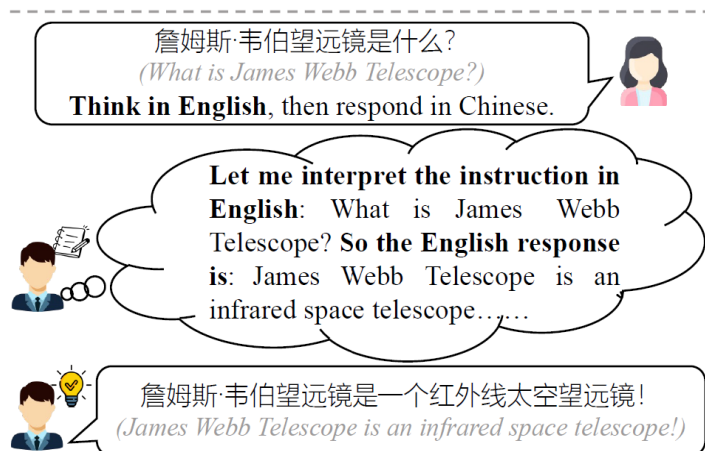


Monolingual QA with Translate-then-Test

Monolingual QA → Translate to English → English Answer → Translate Back to Source Language

Monolingual QA with 'thinking in English' aka

Cross-Lingual Thought Prompting



Model asked to generate intermediate English artifacts

No explicit translation call

The LLM itself is used as a translator **implicitly**

- ✓ Multiple inferences are avoided
- ✓ Input in original language is available to LLM
- ✗ Increased token length for model, Reduces possible input token size

[XLT, PLUG]

Cross-Lingual Thought Prompting (XLT)

Ask the model to implicitly think in a different language

Settings		Reasoning		Understanding		Generation		
		MGSM	XCOPA	XNLI	PAWS-X	MKQA	XL-Sum*	FLORES*
Zero-shot	text-davinci-003							
	Basic Prompt	12.5	70.1	53.3	52.0	29.0	23.7	15.4
	CoT	25.7	70.9	53.0	57.8	30.9	23.8	15.8
	Translate-En	15.7	68.0	54.8	55.0	—	—	—
	XLT	23.9	73.3	62.4	57.1	40.2	25.2	17.7
	gpt-3.5-turbo							
	Basic Prompt	23.3	76.9	52.6	65.5	31.6	24.7	19.1
	CoT	45.5	78.3	54.8	61.0	14.8	25.4	19.7
Few-shot	Translate-En	27.1	75.7	52.2	66.8	—	—	—
	XLT	70.0	80.3	65.5	63.6	42.7	26.1	21.2
	text-davinci-003							
	Basic Prompt	45.5	75.6	59.1	68.7	39.1	26.8	—
	Translate-En	46.5	77.4	56.9	68.5	—	—	—
	XLT	55.4	81.3	67.5	72.2	49.6	27.3	—
	gpt-3.5-turbo							
	Basic Prompt	63.0	80.1	61.4	66.4	43.7	25.5	—
Few-shot	Translate-En	65.1	81.9	58.3	63.7	—	—	—
	XLT	72.5	85.9	65.0	69.1	52.5	27.9	—

Model asked to generate intermediate English artifacts

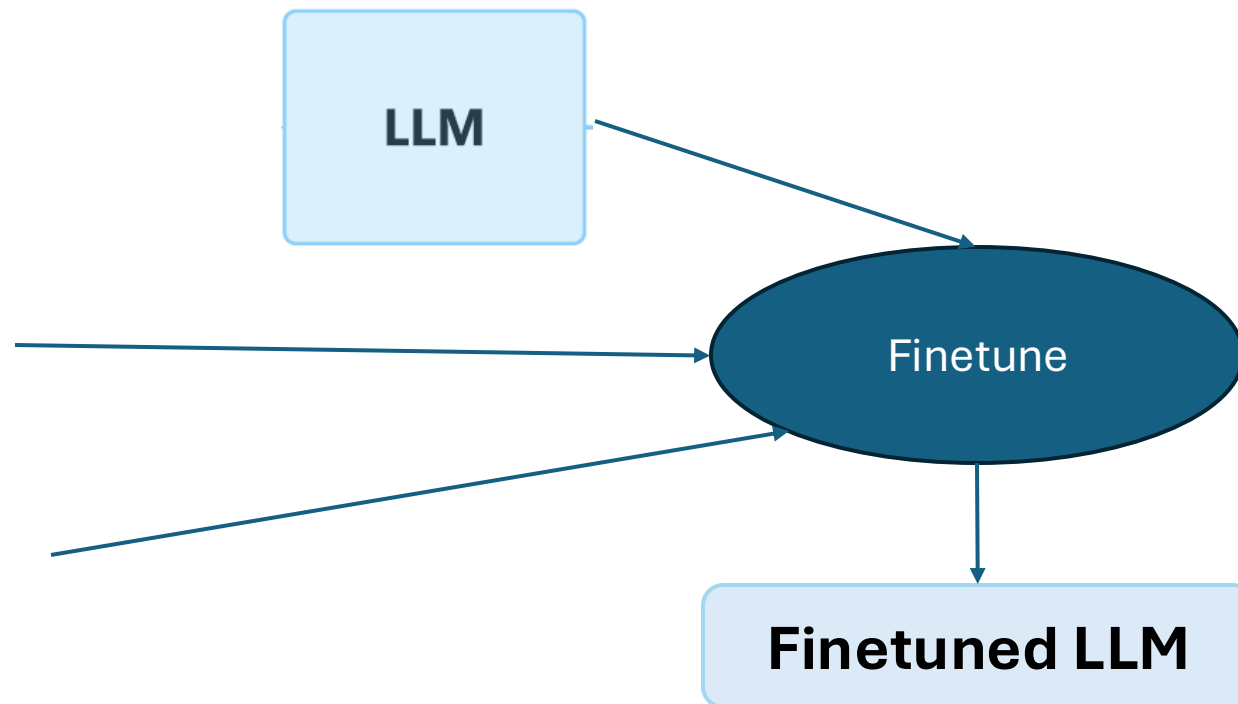
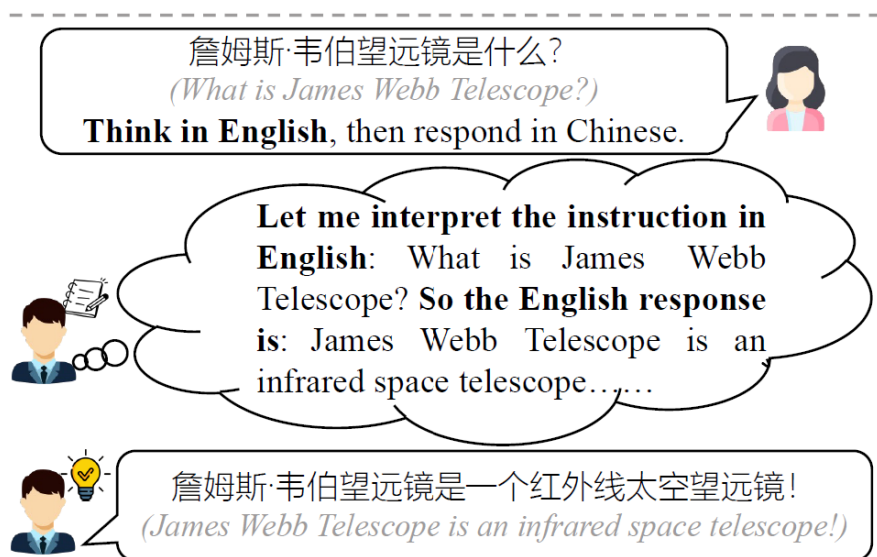
No explicit translation call

The LLM itself is used as a translator **implicitly**

- ✓ Multiple inferences are avoided
- ✓ Input in original language is available to LLM
- ✗ Increased token length for model, Reduces possible input token size

(XLT, PLUG)

TaCo: Instruction tuning with Cross-Lingual Thought data



Limitations

- Reduced maximum sequence length
- Increased latency

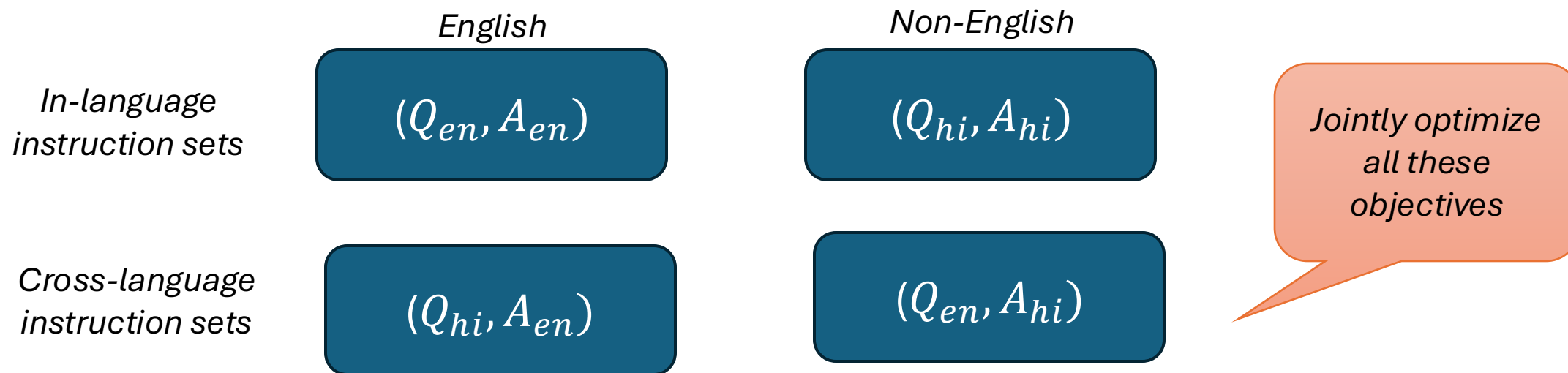
	Standard IFT	TaCo
Nepali	39.3	88.1
Persian	46.5	83.9

Evaluation on translated Vicuna Benchmark with GPT4

The PLUG paper also shows improvements (discussed later)

(TaCo, PLUG)

Cross-Lingual Instruction Tuning



Q_{hi} : मैं अपने समय प्रबंधन कौशल को कैसे सुधार सकता हूँ? मुझे 5 संक्षिप्त बिंदु दीजिये।

A_{en} :

1. Prioritize tasks based on importance and urgency.
2. Set specific goals with deadlines (SMART goals).
3. Use time management tools like calendars and apps.
4. Minimize distractions to maintain focus.
5. Regularly review and adjust your schedule for better efficiency.

Cross-lingual instruction help the model to align languages and think in English

Avoids intermediate English generation of Cross-lingual Thought Prompting

[SDRRL,XCOT,CIT]

Code-switched Instruction Tuning

Code-switching: alternating between two or more languages in conversation

Generate code-switched instruction by replacing some source words with their target language translations

Q_{hi} : मैं अपने time प्रबंधन skill को कैसे सुधार सकता हूँ? मुझे 5 brief बिंदु दीजिये।

A_{en} :

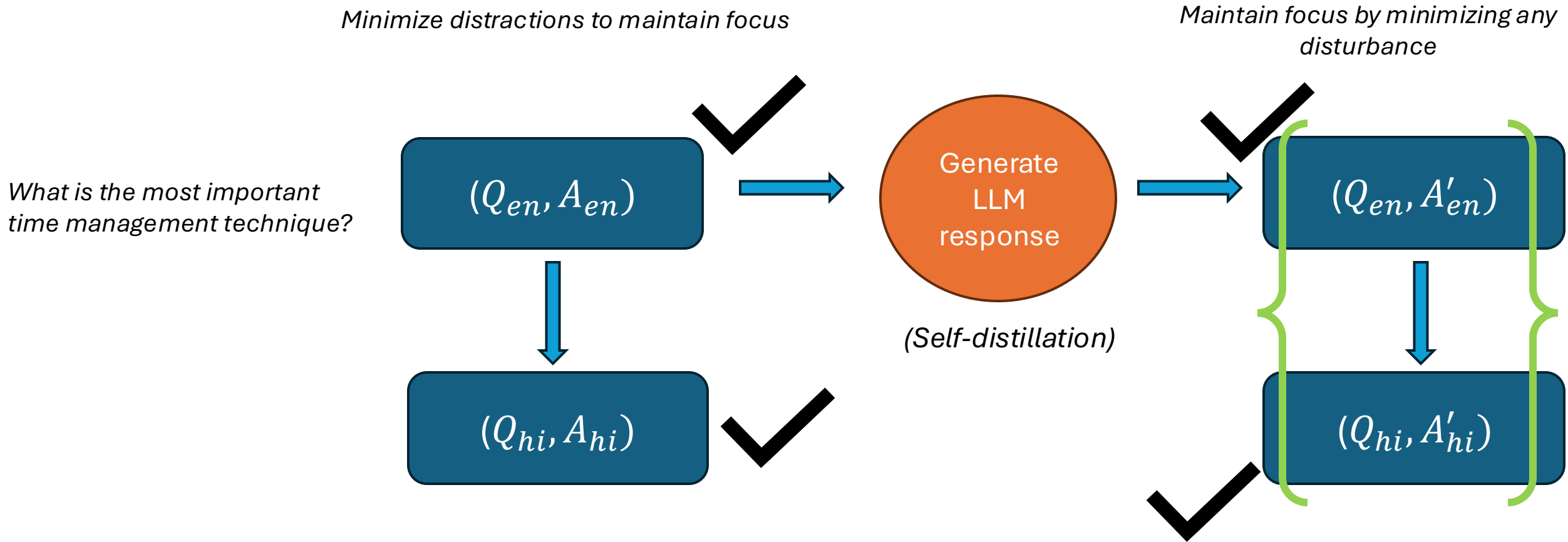
1. महत्व और तात्कालिकता के आधार पर कार्यों को प्राथमिकता दें।
2. समय सीमा (स्मार्ट लक्ष्य) के साथ विशिष्ट लक्ष्य निर्धारित करें।
3. कैलेंडर और ऐप्स जैसे समय प्रबंधन टूल का उपयोग करें।
4. फोकस बनाए रखने के लिए विकर्षणों को कम करें।
5. बेहतर दक्षता के लिए नियमित रूप से अपने शेड्यूल की समीक्षा करें और उसे समायोजित करें।

Code-switching help the model to align languages better at sub-sentence level

Some optimizations to improve models further

- Retain English instructions, code-switch only the inputs and outputs
- Do code-switching during pre-training as well

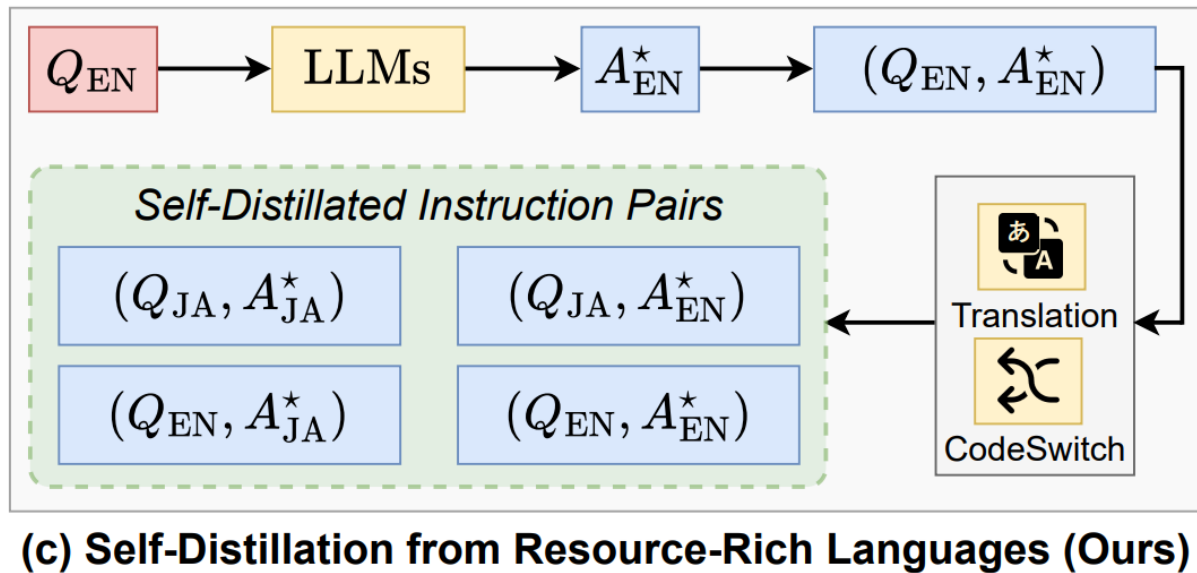
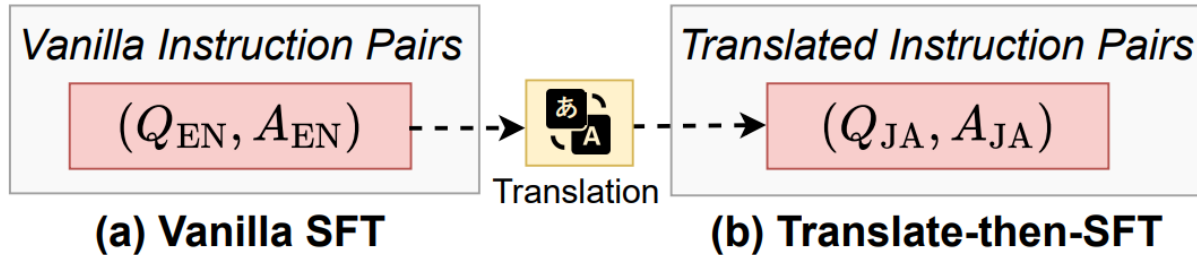
Self-Distillation from Resource Rich Language



Using the model's own responses can help improving the model's own representation space

Putting it all together

[SDRRL]



+

Cross-lingual Instructions

Translation Task

Sentence Completion Task

(these tasks are added so model preserves native language competency)

Summary Results

[SDRRL]

	BELE.	XL-SUM	FLORES	MKQA	AVG.
<i>Performance on Target Language</i>					
SFT	42.24	<u>16.48</u>	18.45	38.86	29.01
T-SFT	<u>42.77</u>	15.32	16.59	43.40	29.52
CIT	42.53	15.75	<u>20.49</u>	<u>43.70</u>	<u>30.62</u>
XCOT	41.19	15.79	17.21	42.04	29.06
SDRRL	43.67	17.89	25.86	44.63	33.01
<i>Performance on English Language</i>					
SFT	<u>60.19</u>	15.25	<u>28.49</u>	<u>39.62</u>	<u>35.89</u>
T-SFT	58.70	<u>15.63</u>	23.72	37.43	33.87
CIT	58.66	15.42	18.31	36.67	32.27
XCOT	57.73	14.90	23.96	37.94	33.63
SDRRL	60.67	16.24	29.47	40.32	36.68

SFT: FT on English data

T-SFT: source and target translated

CIT: target translated

XCOT: source translated + source code-switching

(This summary is for the SeaLLM backbone LLM, results in main paper are for LLama)

Bringing together all these objectives and data augmentations:

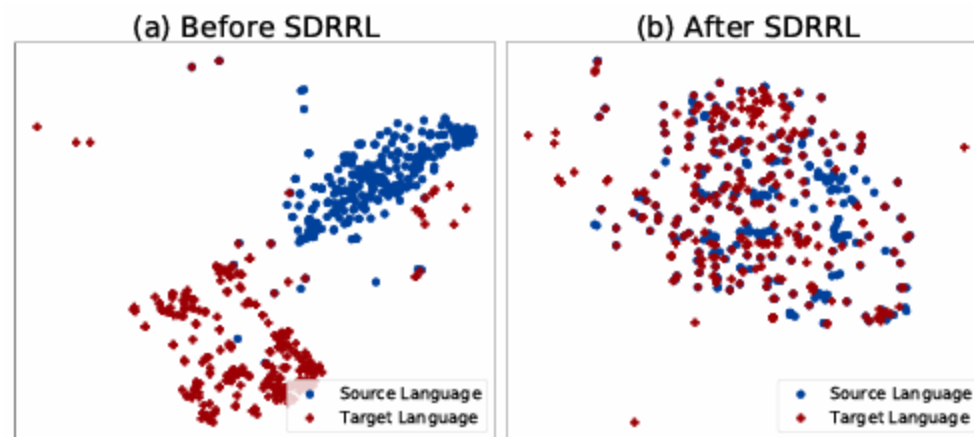
- **Helps improve overall response quality across multiple tasks**
- **Retains English performance**

IFT Data Transformation Helps [SDRRL]

		NLU Avg.		NLG Avg.	
		TAR.	ENG	TAR.	ENG
1	Full Method	50.58	66.29	28.24	31.69
2	- $\mathcal{D}_{TL}$ and $\mathcal{D}_{LT}$	49.56	65.93	26.15	30.55
3	- $\mathcal{D}_{synth} + \mathcal{D}$	48.59	65.10	25.16	30.10
4	- $\mathcal{D}_{mt}$ and $\mathcal{D}_{comp}$	<u>50.41</u>	<u>66.01</u>	26.61	30.19
5	- Code Switching	50.37	65.94	<u>27.13</u>	<u>30.69</u>
6	Only $\mathcal{D}_{mt}$ and $\mathcal{D}_{comp}$	41.25	61.61	17.89	22.28

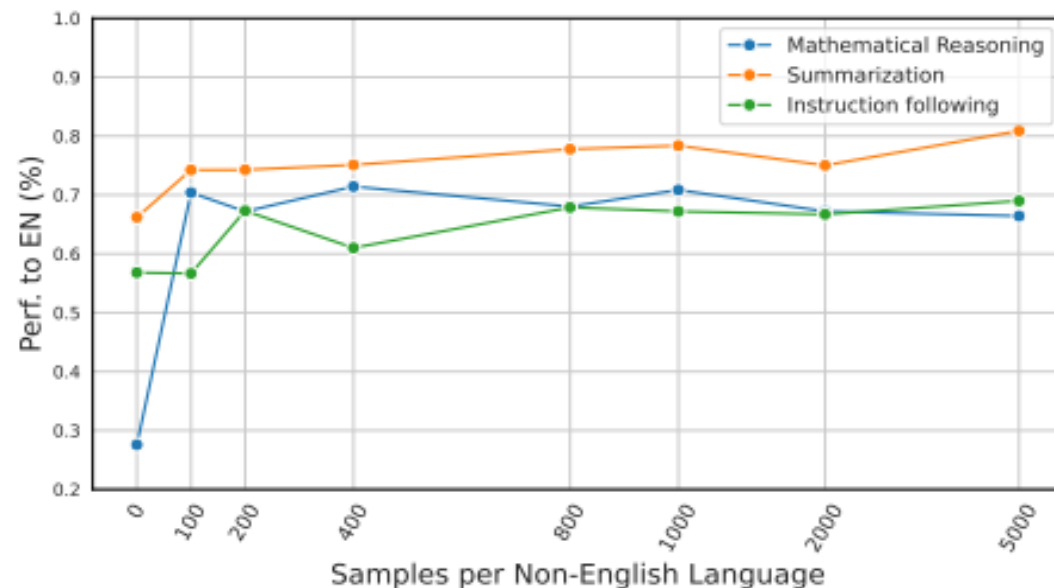
Table 6: Ablation study. Average scores of target language (TAR.) and English (ENG) on natural language understanding task (NLU, including BELEBELE) and natural language generation tasks (NLG, including FLORES, XL-SUM ROUGE-L, and MKQA) are reported.

- Using the LLMs own responses is a very useful method to improve cross-lingual transfer
- The MT and sentence completion tasks are very useful
- The cross-lingual instruction tuning tasks are also complementary
- Code-switching (on input side) has modest benefits



How much non-English IFT Data?

[MultilingualPostTrainer]



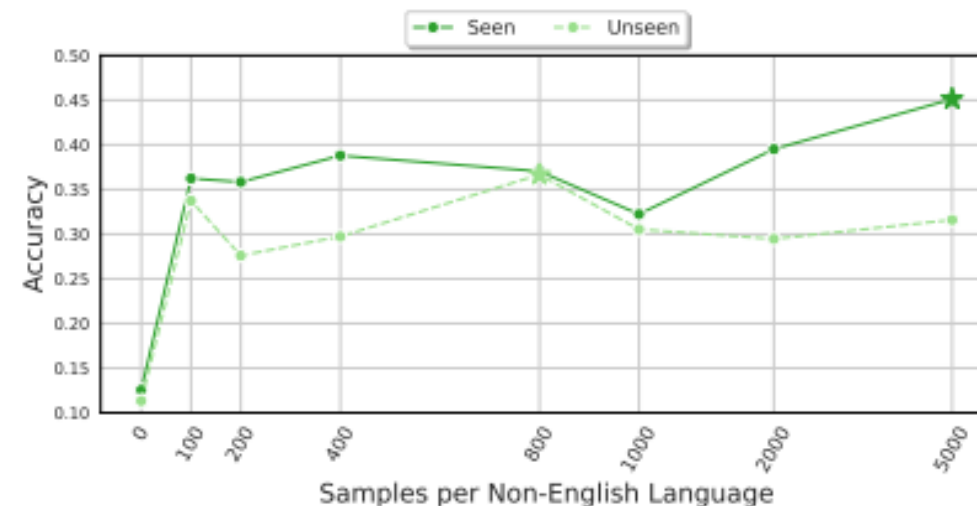
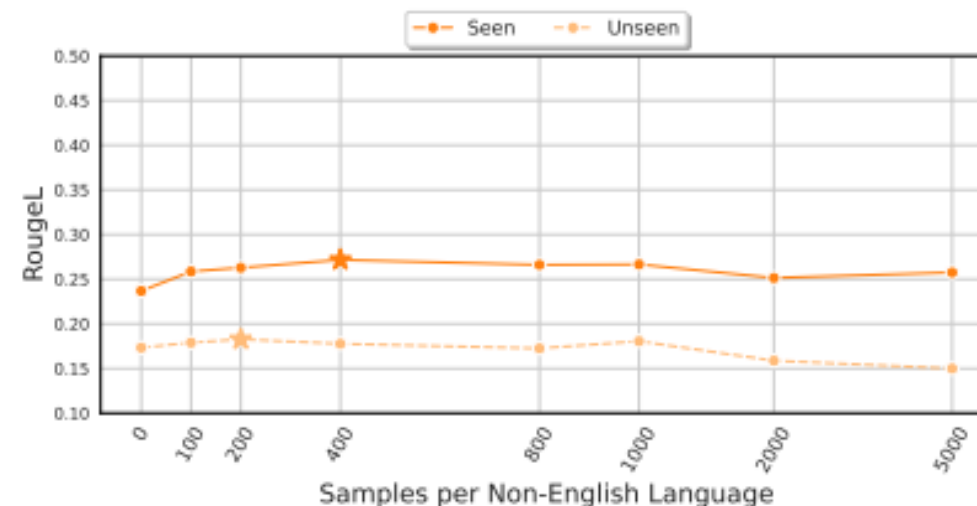
As we increase number of non-English language samples:

- Small models: Multilingual performance improvements are task-dependent
- Large models: Multilingual performance plateaus after the addition of 200– 400 samples per language for each task.

Seen vs Unseen language

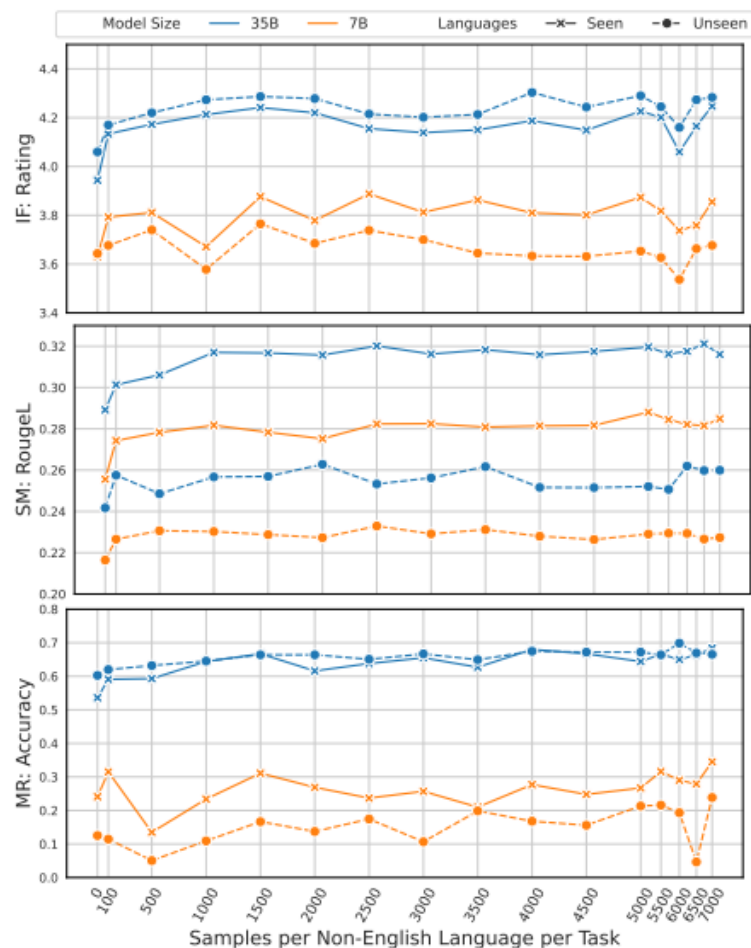
[MultilingualPostTrainer]

- Unseen languages benefit from improvements on seen languages
- But suffer if seen language performance deteriorates
- This means that the performance trend on the unseen language is dependent on the seen language performance trend



Multi-task Training Benefits

[MultilingualPostTrainer]

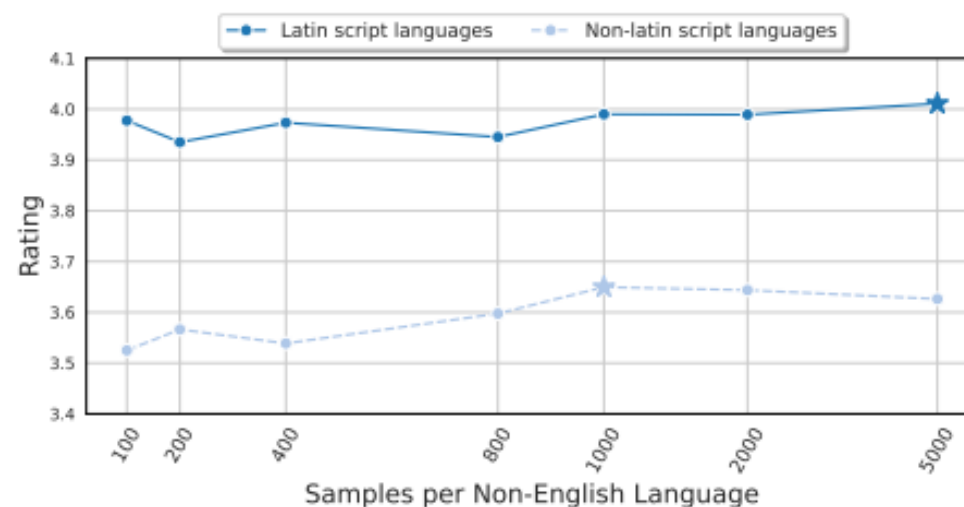
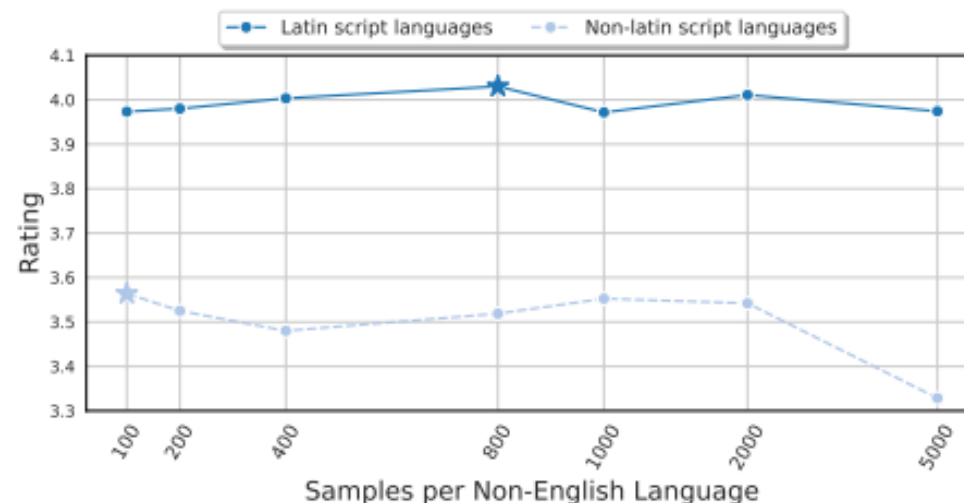


- Non-English languages benefit from task diversity in large models.
- With sufficient scale, models are able to transfer to unseen languages for some tasks.

Script vs Language

- Cross-lingual transfer from Latin script languages insufficient for non-Latin script languages – Script dependency of cross-lingual transfer
- Mathematical Reasoning may benefit from Latin-script data as it is not as linguistically motivated as IF and summarization tasks.

[MultilingualPostTrainer)]

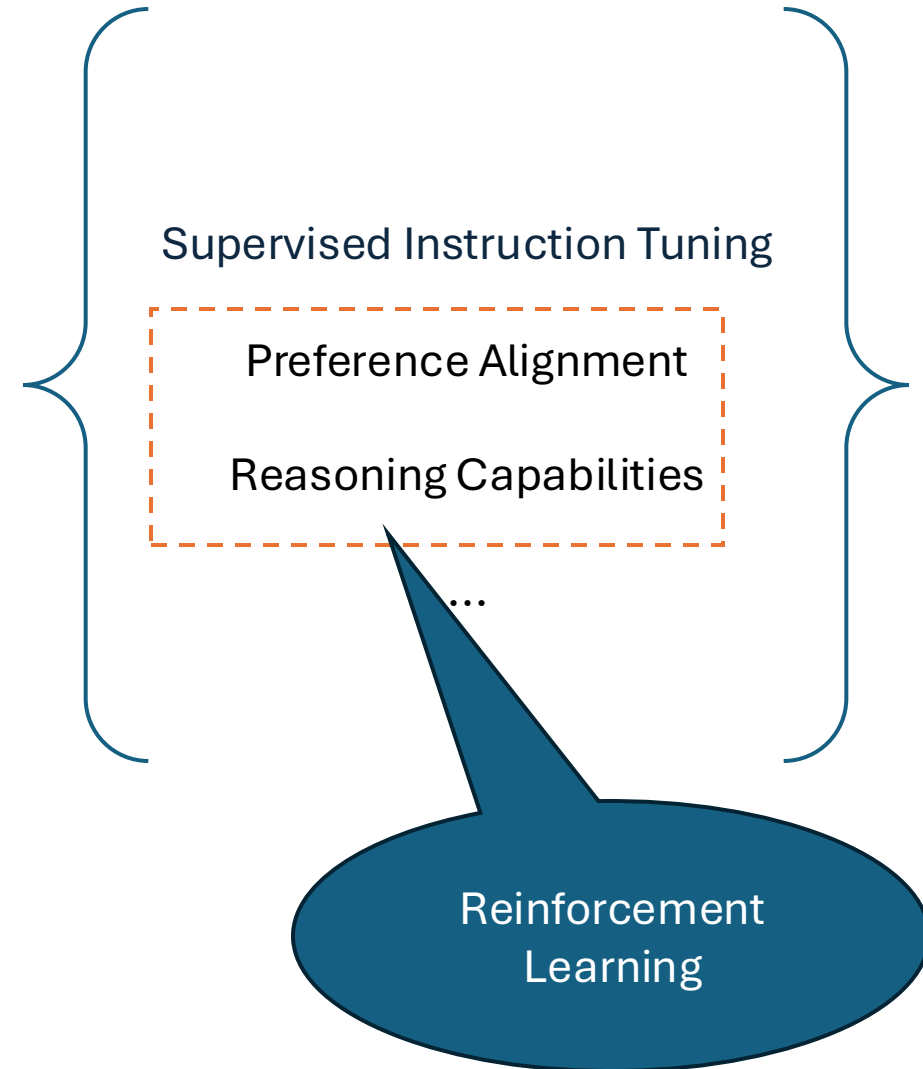
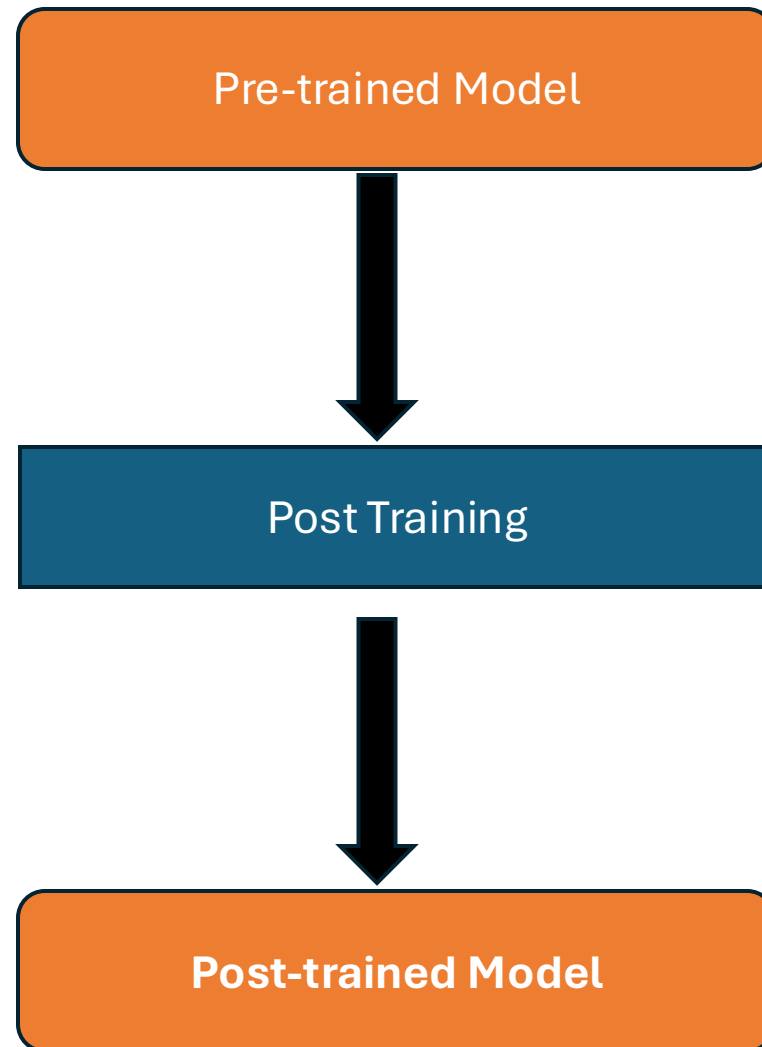


Summary and Recommendations

- Machine Translation is a **good starting point** to create IFT data
- **Use LLMs to generate culture/region-specific data** in English/target-language
- Improve alignment between English and other languages using methods like **cross-lingual instruction tuning, romanized/code-switched data**
- **Machine Translation is an important task** for cross-lingual alignment

Beyond Multilingual IFT: What's Next?

- Multilingual IFT gives us *instruction-following ability across languages*
- But two core challenges remain:
 - **Preference Alignment:** Responses may still be not well-calibrated/nuanced, unsafe, biased, or culturally misaligned across languages
 - **Reasoning:** Models often fail at multi-step thinking, planning, factual consistency

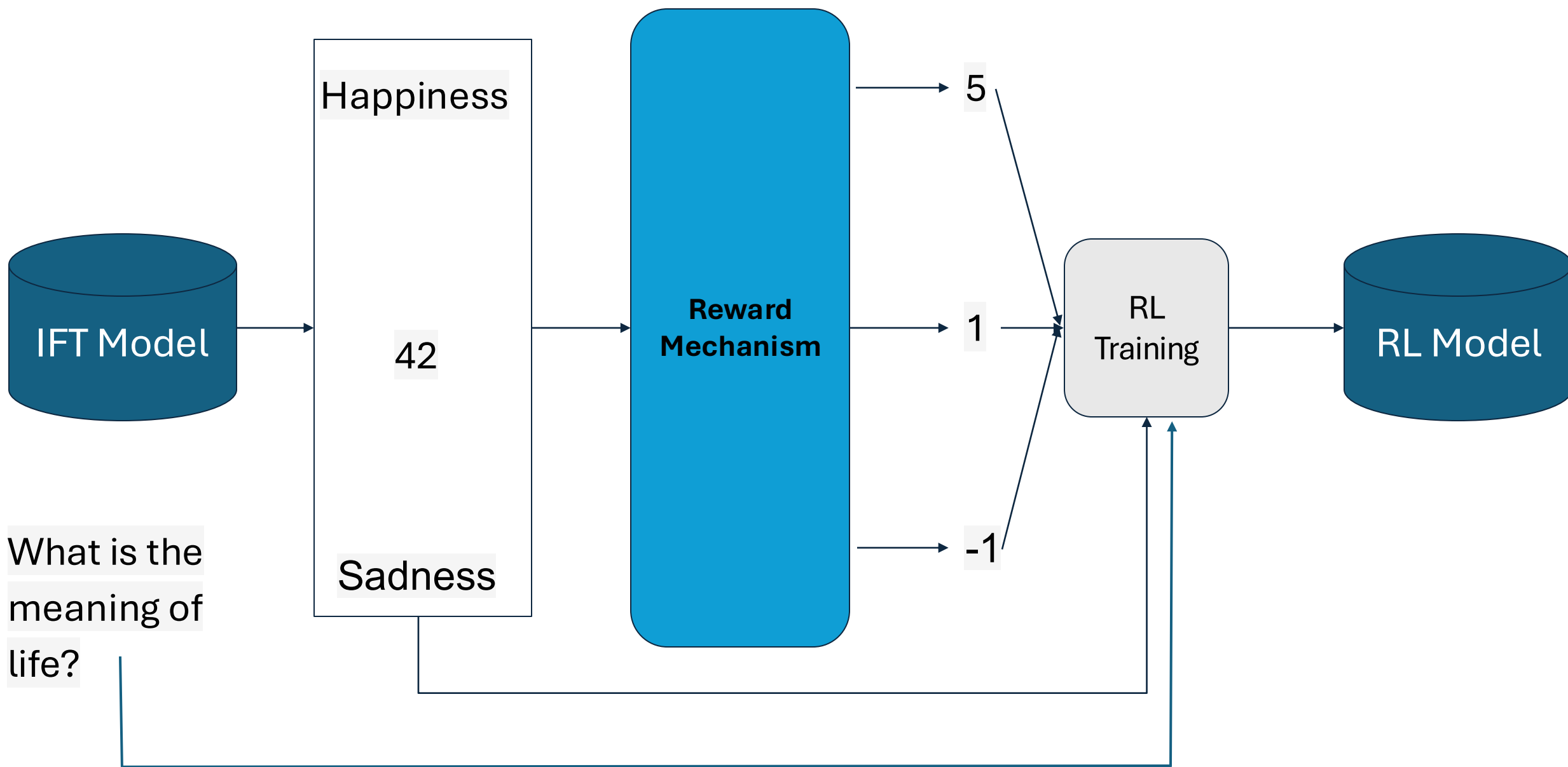


Reinforcement Learning

Tweaking model behavior

Why RL?

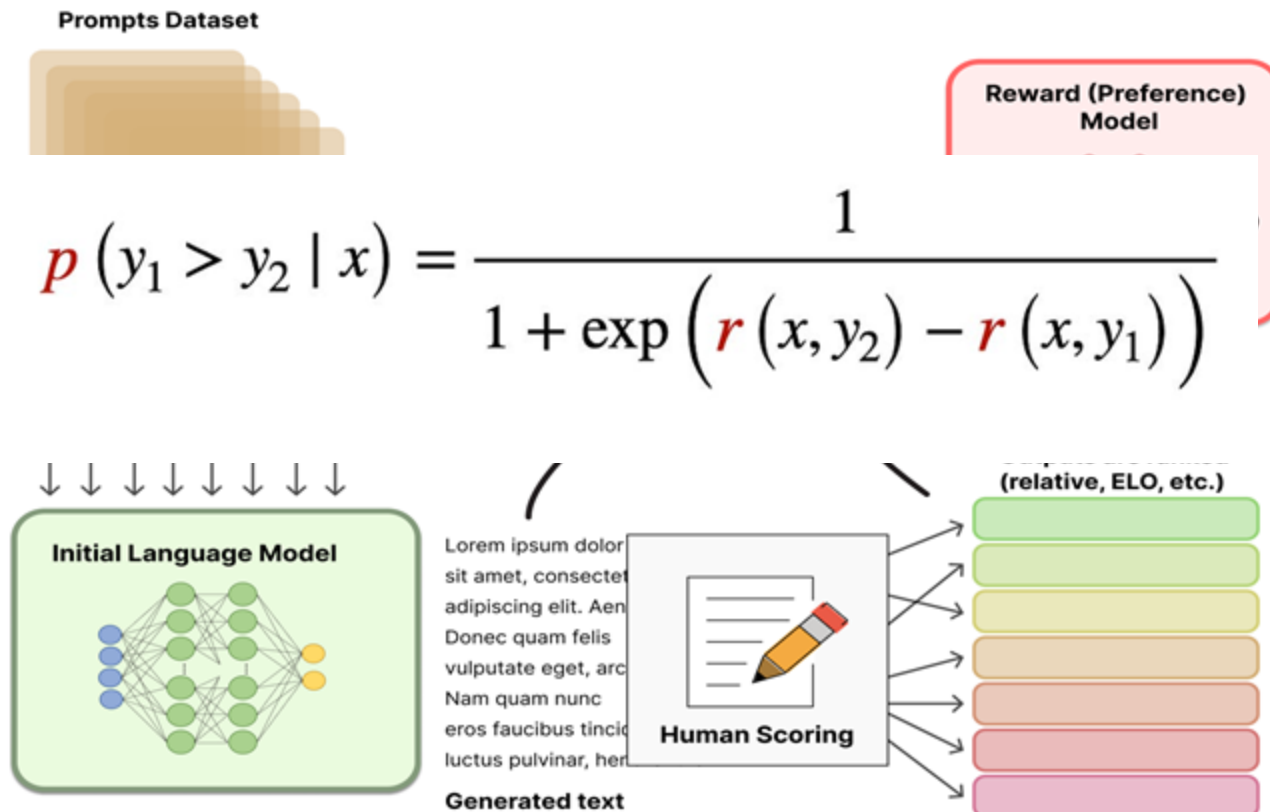
- To encourage model to explore CoT without explicit supervision
- To avoid needing well written responses for supervised training
 - Saves massive human effort
- To incorporate human feedback at a high level
- To get to ASI/AGI



Reward Mechanisms (Broadly)

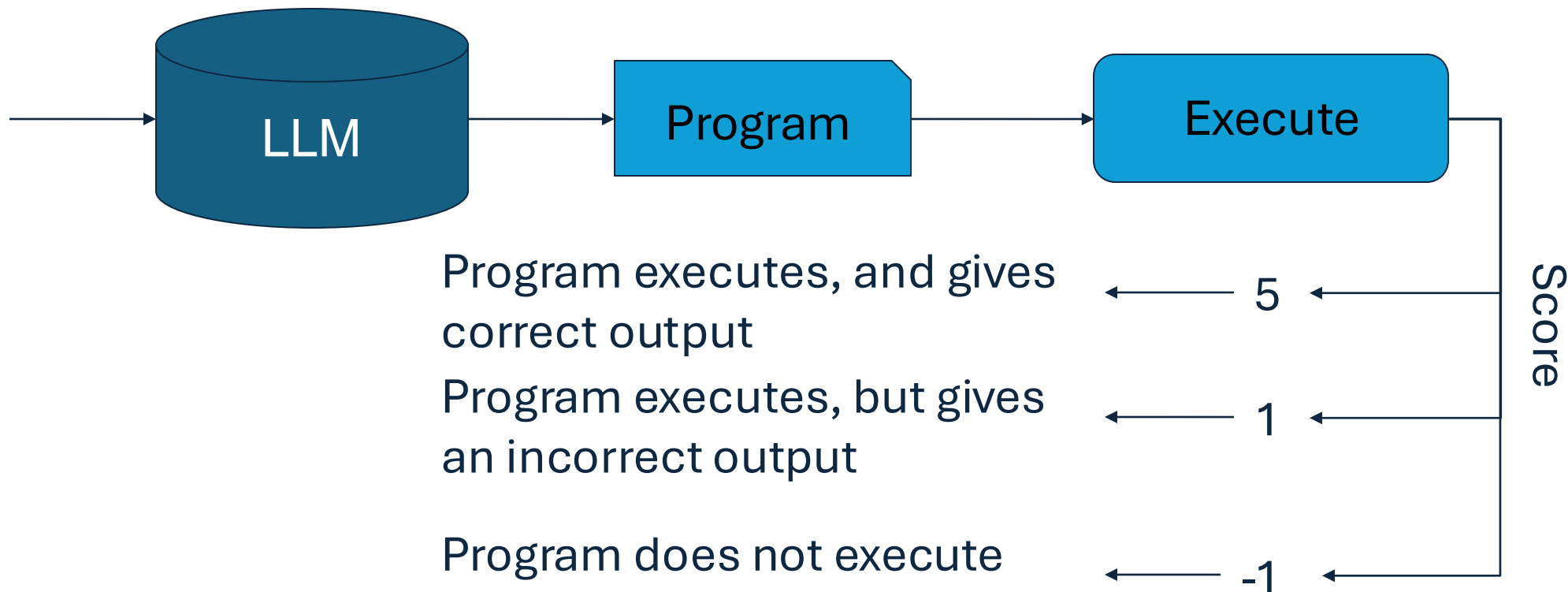
- Via human signals (RLHF)
 - Rewards from a model based on human feedback on model responses
- Via Non-Human signals
 - Via machine feedback (RLMF)
 - Rewards from an execution of a program
 - Via Grading LLM or AI models (RLGF/RLAIF)
 - Rewards from a LLM or some AI model

Building Reward Models From Human Feedback

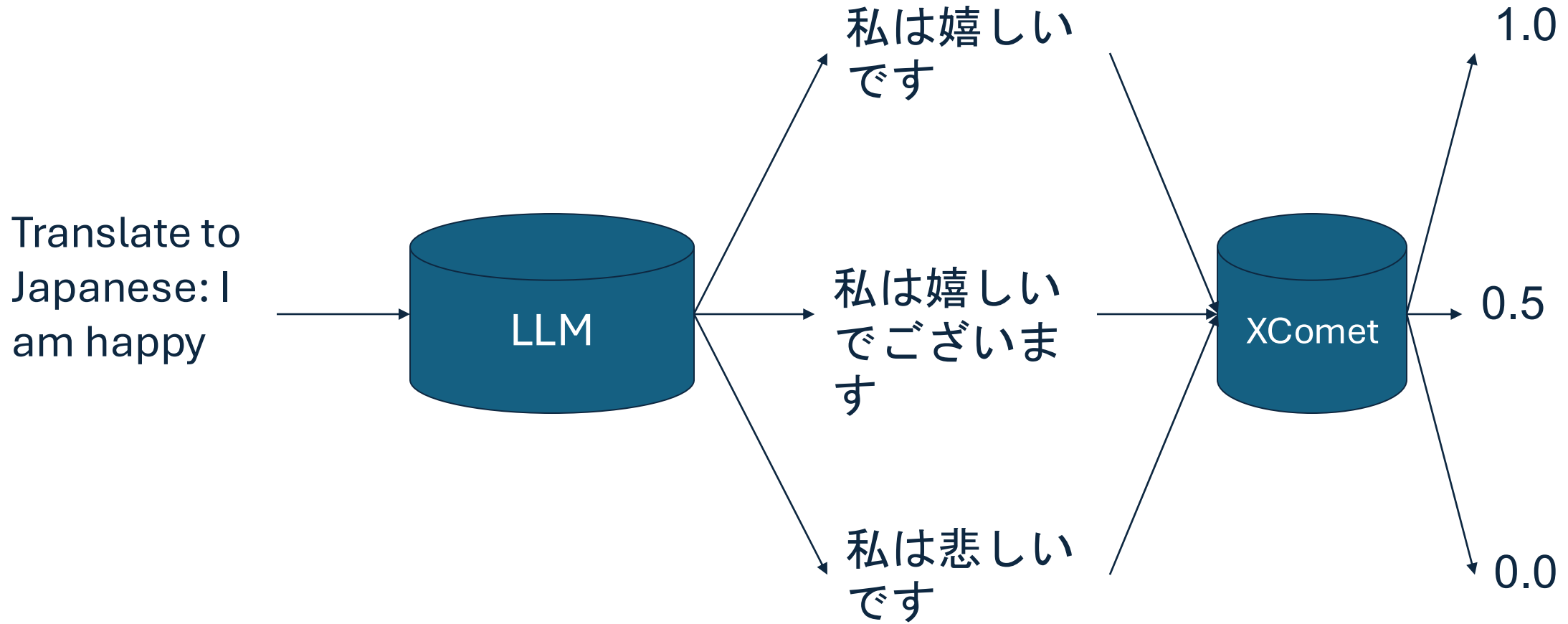


Rewards From Machine Feedback (RLMF)

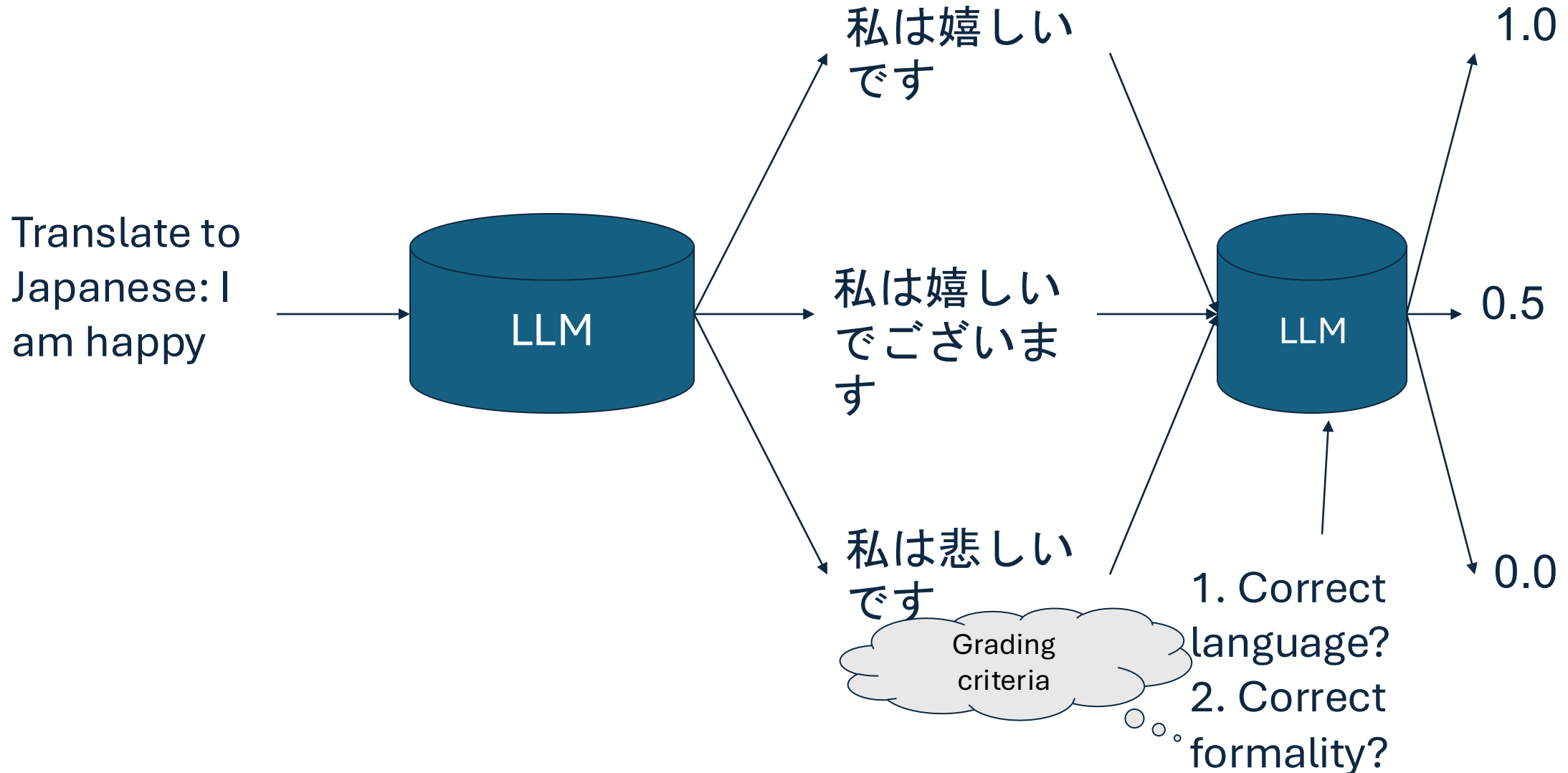
Write a
program to
add two
numbers

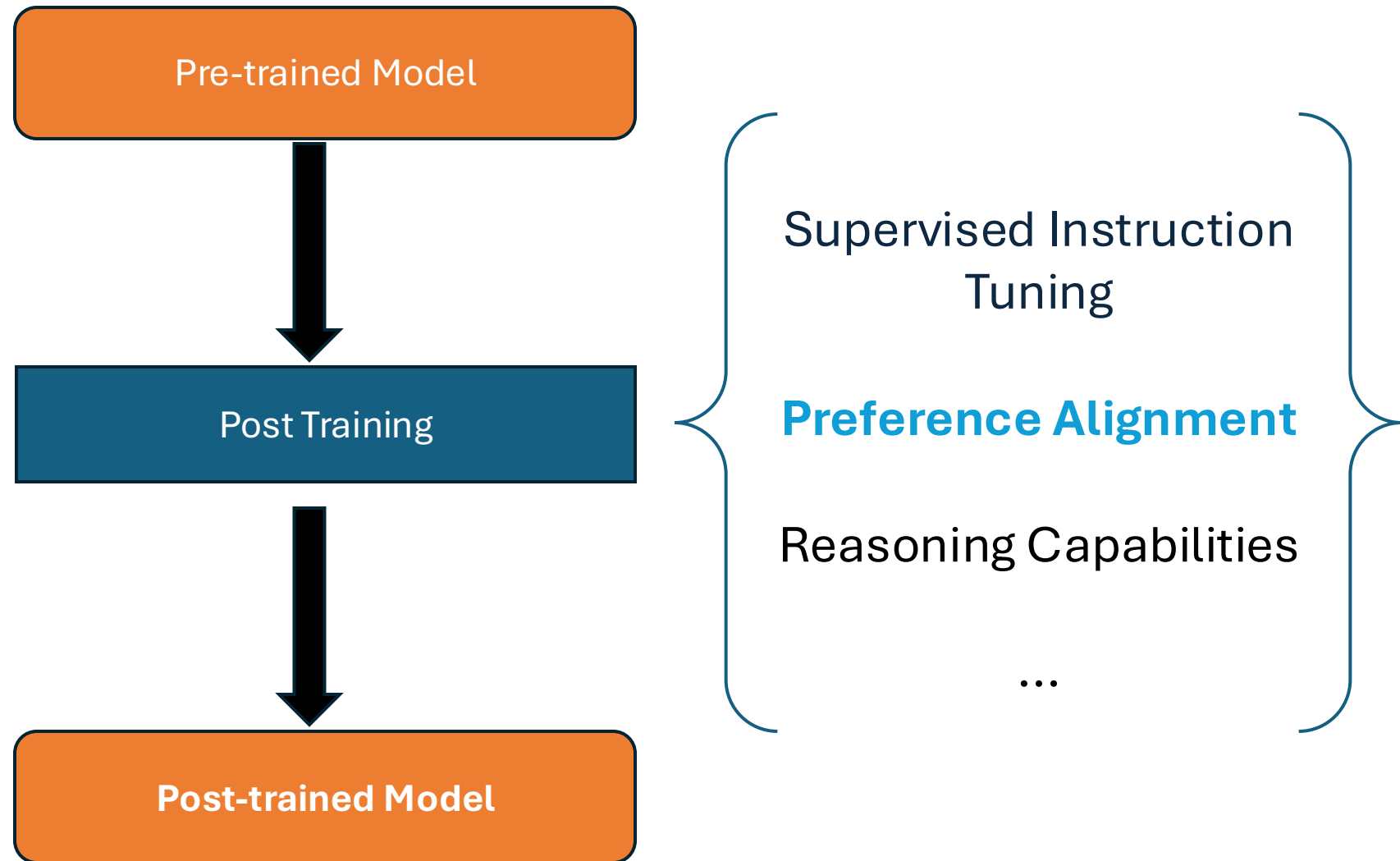


Rewards From AI Feedback (RLAIF)



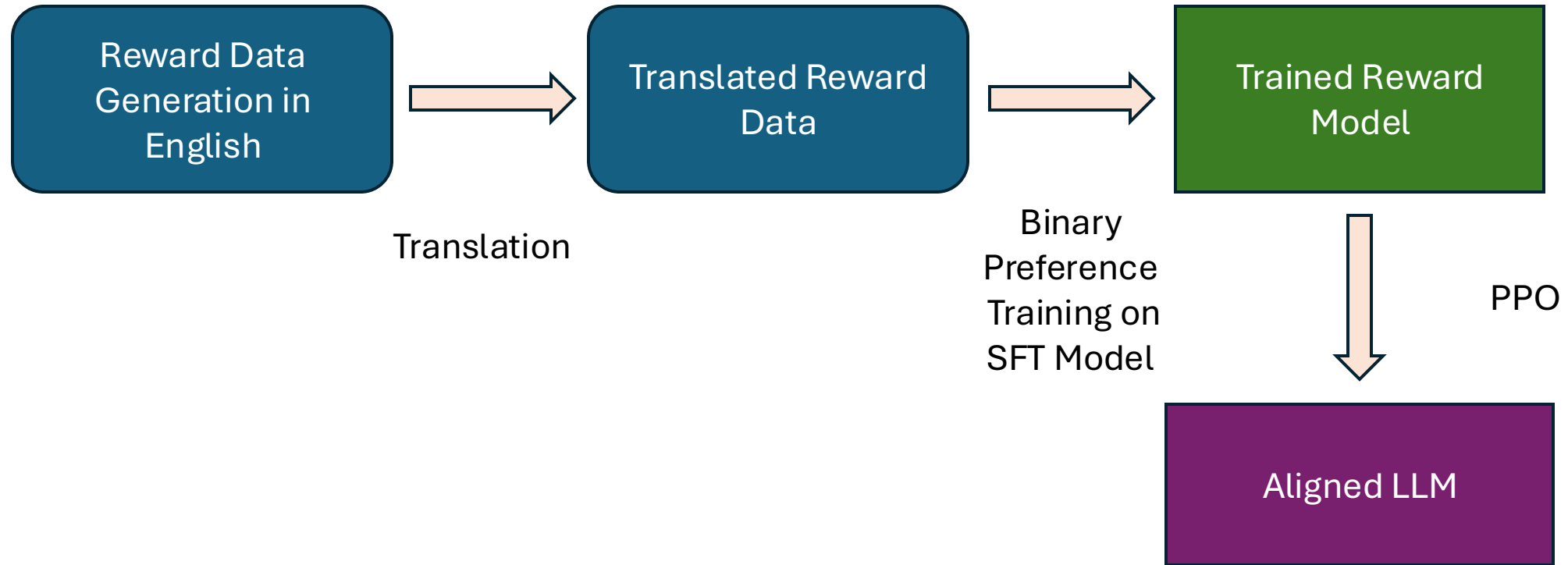
Rewards From Grading LLM (RLGF)



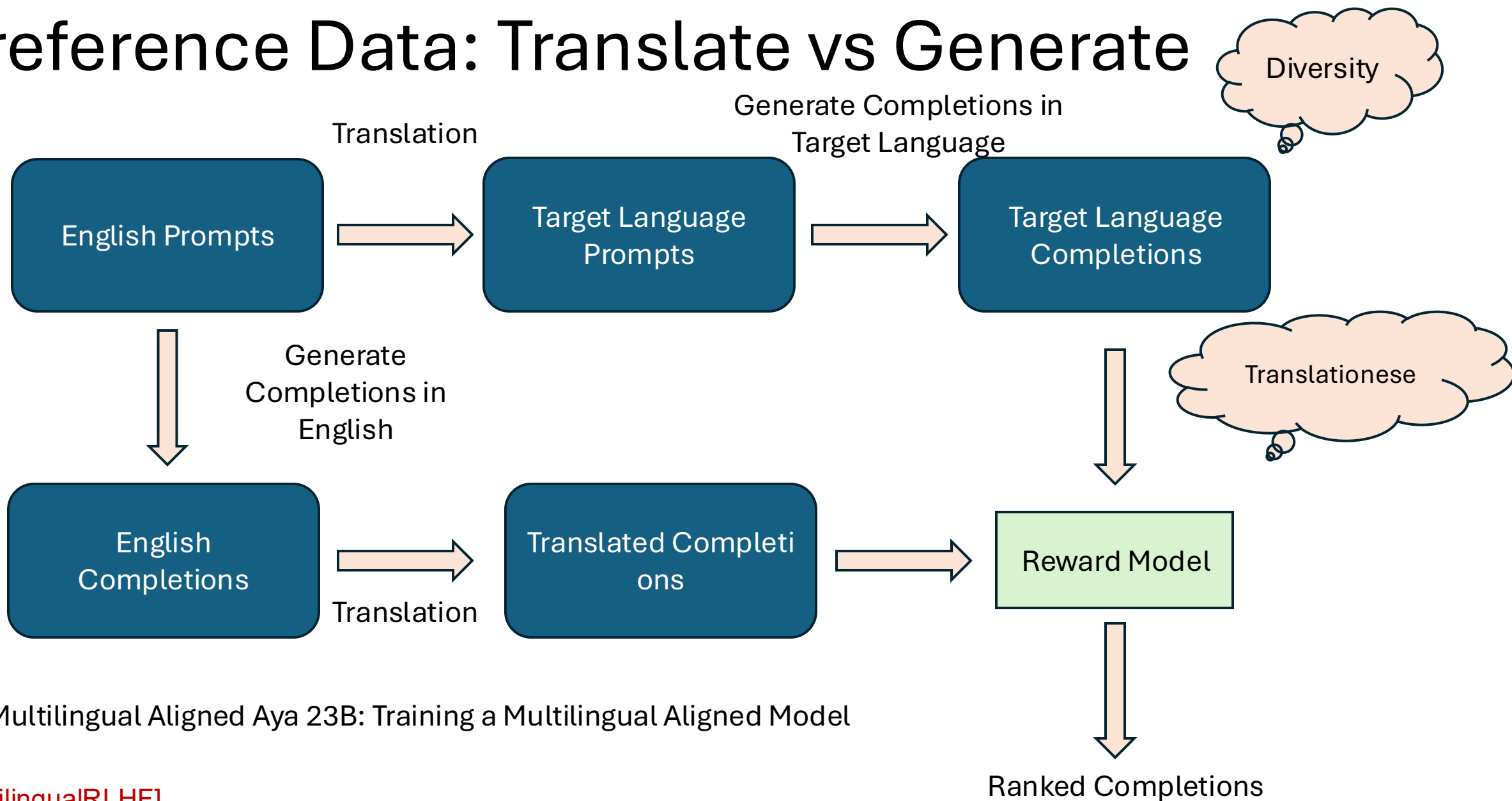


Multilingual Preference Alignment

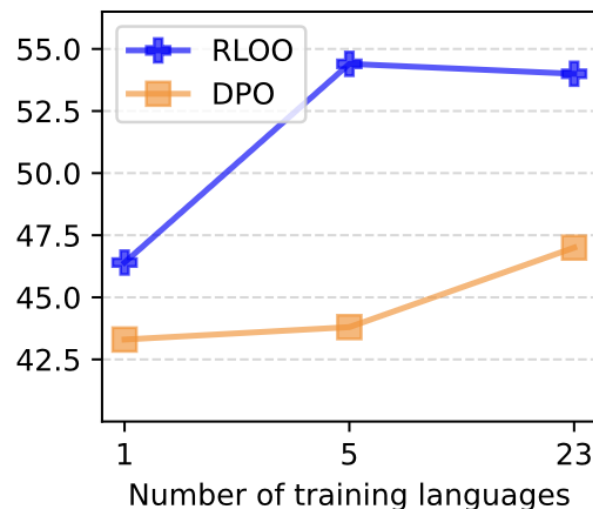
Target Language Reward Model



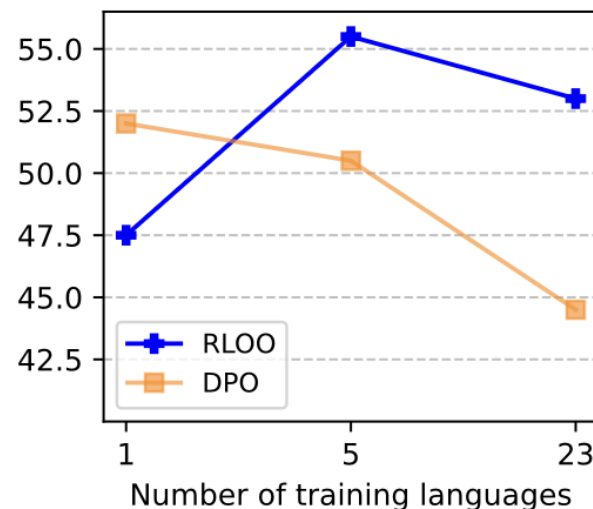
Preference Data: Translate vs Generate



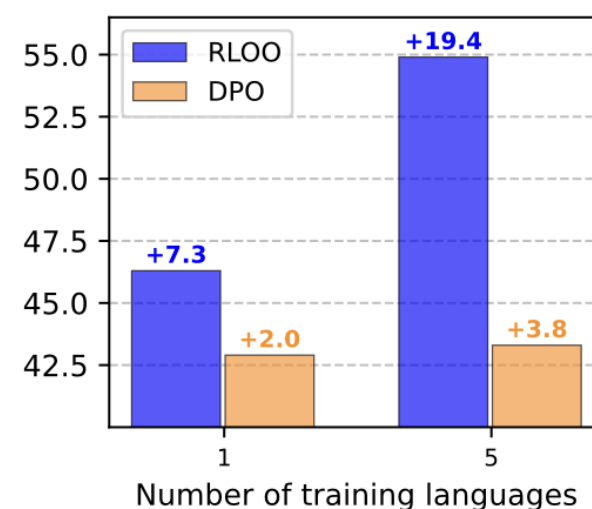
Impact of Number of Languages during RL



(a) Avg. win% on **23** languages



(b) Win% on **English**



(c) Avg. win% on **unseen** langs.

- Preference optimization using English helps
- More languages leads to further increase
- Unseen languages benefits from more languages
- RLOO is better than DPO

Impact of Number of Training Instances during RL

		Win%	English Loss%	$\Delta W-L\%$
DPO	EN-1	52.0	33.5	18.5
	ML-5	50.5	28.5	22.0
	ML-23	44.5	36.5	8.0
	ML-23*	57.5	31.0	26.5
RLOO	EN-1	47.5	38.5	9.0
	ML-5	55.5	30.5	25.0
	ML-23	53.0	37.0	16.0
	ML-23*	53.0	35.0	18.0

		Average 23 Languages		
		Win%	Loss%	$\Delta W-L\%$
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

- More instances (ML-23* vs ML-23) leads to improvements for DPO and RLOO

Can English Reward Models be used in Cross-lingual Setting?

RewardBench	Category	LLAMA-3.2-3B-IT					QWEN2.5-3B-IT				
		Chat	Chat(H)	Safety	Reason	Avg.	Chat	Chat(H)	Safety	Reason	Avg.
SPANISH	Target	79.1	67.3	88.0	65.5	75.0	80.7	68.2	84.8	68.2	75.5
	English	86.3	69.3	89.3	72.4	79.3	82.7	68.0	88.3	73.6	78.1
	Δ	+7.2	+2.0	+1.3	+6.9	+4.3	+2.0	-0.2	+3.5	+5.4	+2.6
ITALIAN	Target	75.4	62.5	88.5	65.7	73.0	77.1	67.8	85.7	72.8	75.8
	English	83.0	69.3	88.7	75.1	79.0	83.2	68.2	88.4	76.0	79.0
	Δ	+7.6	+6.8	+0.2	+9.4	+6.0	+6.1	+0.4	+2.7	+3.2	+3.2
KOREAN	Target	69.6	58.8	80.9	60.1	67.3	68.4	63.2	80.9	61.4	68.5
	English	69.8	59.4	84.3	73.0	71.6	70.7	61.6	85.4	73.6	72.8
	Δ	+0.2	+0.6	+3.4	+12.9	+4.3	+2.3	-1.6	+4.5	+12.2	+4.3
CHINESE	Target	68.7	59.9	81.2	52.6	65.6	69.8	64.7	81.8	61.3	69.4
	English	54.7	64.0	82.6	79.3	70.2	58.7	67.8	84.3	78.2	72.2
	Δ	-14.0	+4.1	+1.4	+26.7	+4.6	-11.1	+3.1	+2.5	+16.9	+2.8

English RMs show larger gains compared to target language RMs

Table 1: Multilingual RewardBench evaluation results on the target language ("Target") and English ("English") RMs. " Δ " denotes the accuracy gain of English RMs compared to the target language RMs. English RMs show higher average scores in the lingual axis than target language RMs. Also, English RMs excel target language RMs in reasoning ("Reason") tasks with diverse evaluation sub-categories.

English vs non-English Reward Models for Alignment

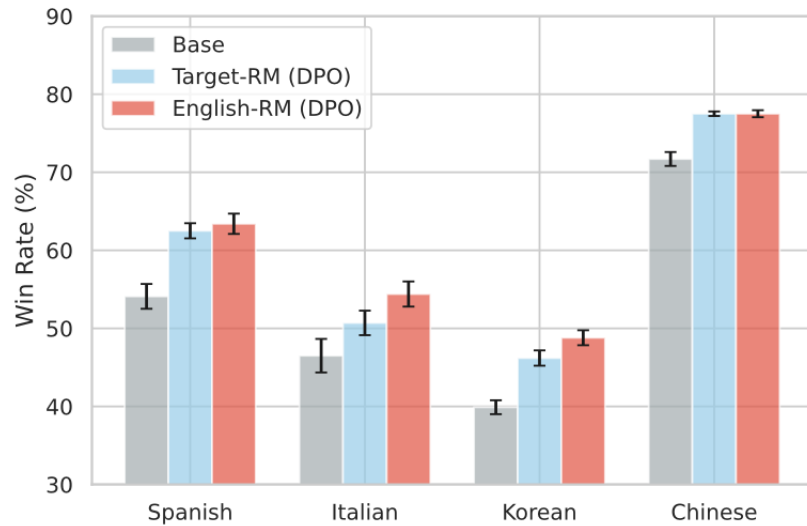


Figure 3: Multilingual AlpacaEval results of Qwen2.5-7B-Instruct models fine-tuned with DPO on on-policy generations for four non-English languages over fine runs. The alignment data were labeled with either English RM or target language RM. Results are averaged over 5 runs.

- English RM --> improvement across languages --> beats target language RM (translated)
- English being the predominant language?
- High-Quality English preference data is a good choice and is cost-effective
- RMs based on MLM objective are better for a good XL reward signal
- Lesson: Convert your LLMs to MLMs before reward modeling
 - MNTP objective

Impact of Cross-lingual Preference Dataset

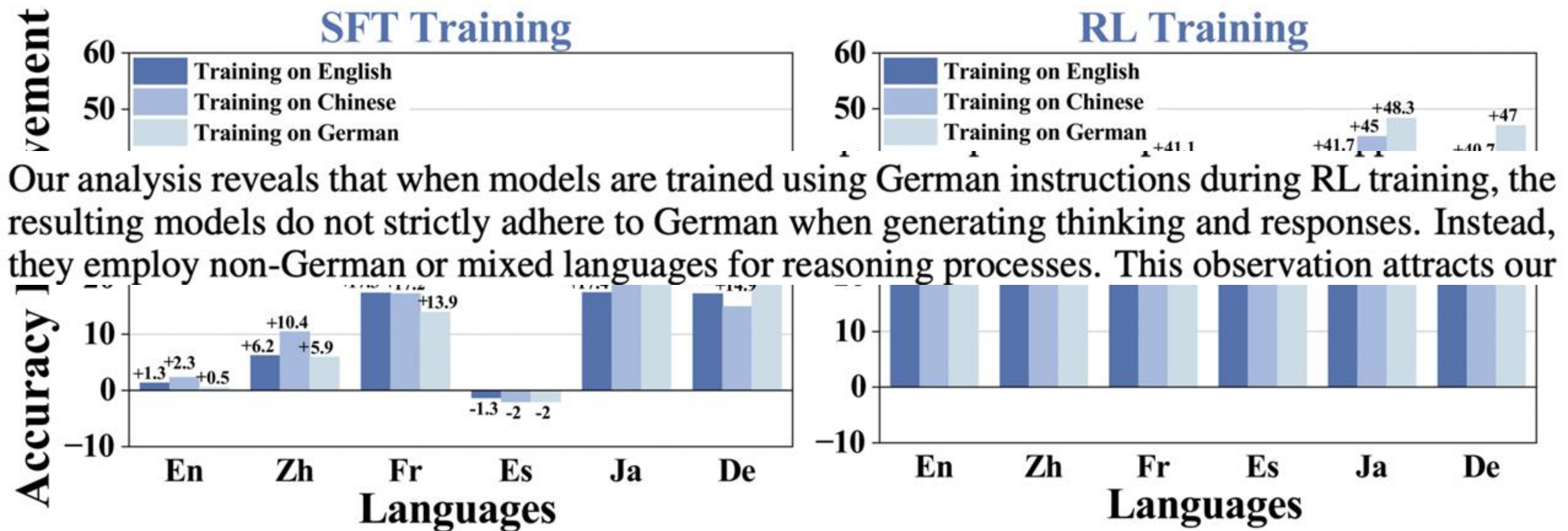
	Low		High	
	mono	cross	mono	cross
PAWS-X	-	-	58.43	61.94
XCOPA	47.26	49.71	52.15	54.68
Self-Instruct*	3.25	9.16	12.14	14.71
XL-Sum	3.38	9.99	12.52	16.61
FLORES(f)	0.85	2.89	4.57	9.07
FLORES(t)	1.55	5.64	8.45	16.98

Table 4: An ablation study of xLLMs-100 using monolingual and cross-lingual human feedback data on low- and high-resource languages.

[xLLMs]

- Cross-lingual feedback data, leads to better generalization compared to monolingual feedback data
- Gains on low-resource languages is more compared to high-resource languages

RL is effective for XL transfer than SFT



Is LLM-as-judge reliable for all languages?

[MultilingualJudgeReliability]

**Good but not
consistent!**

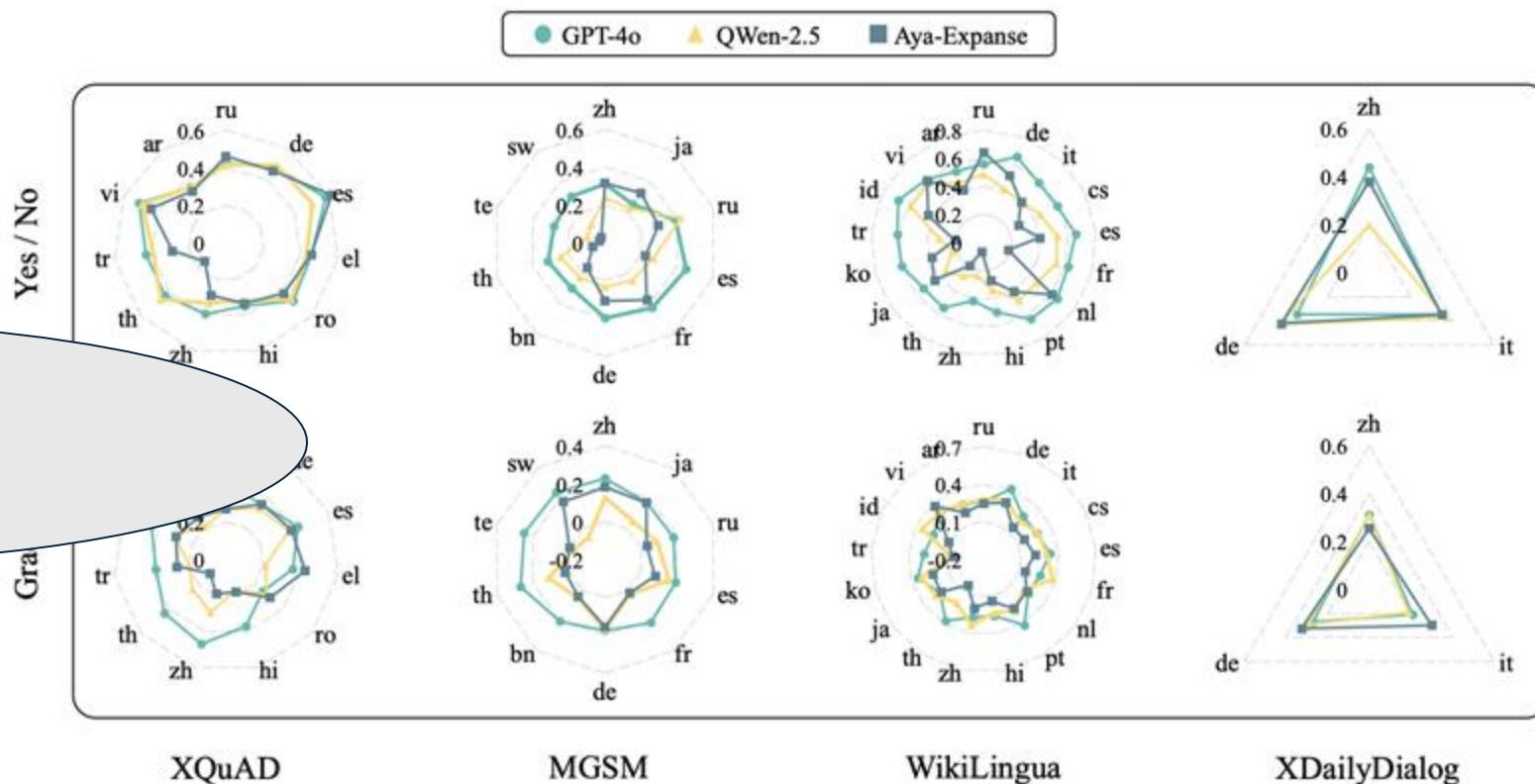
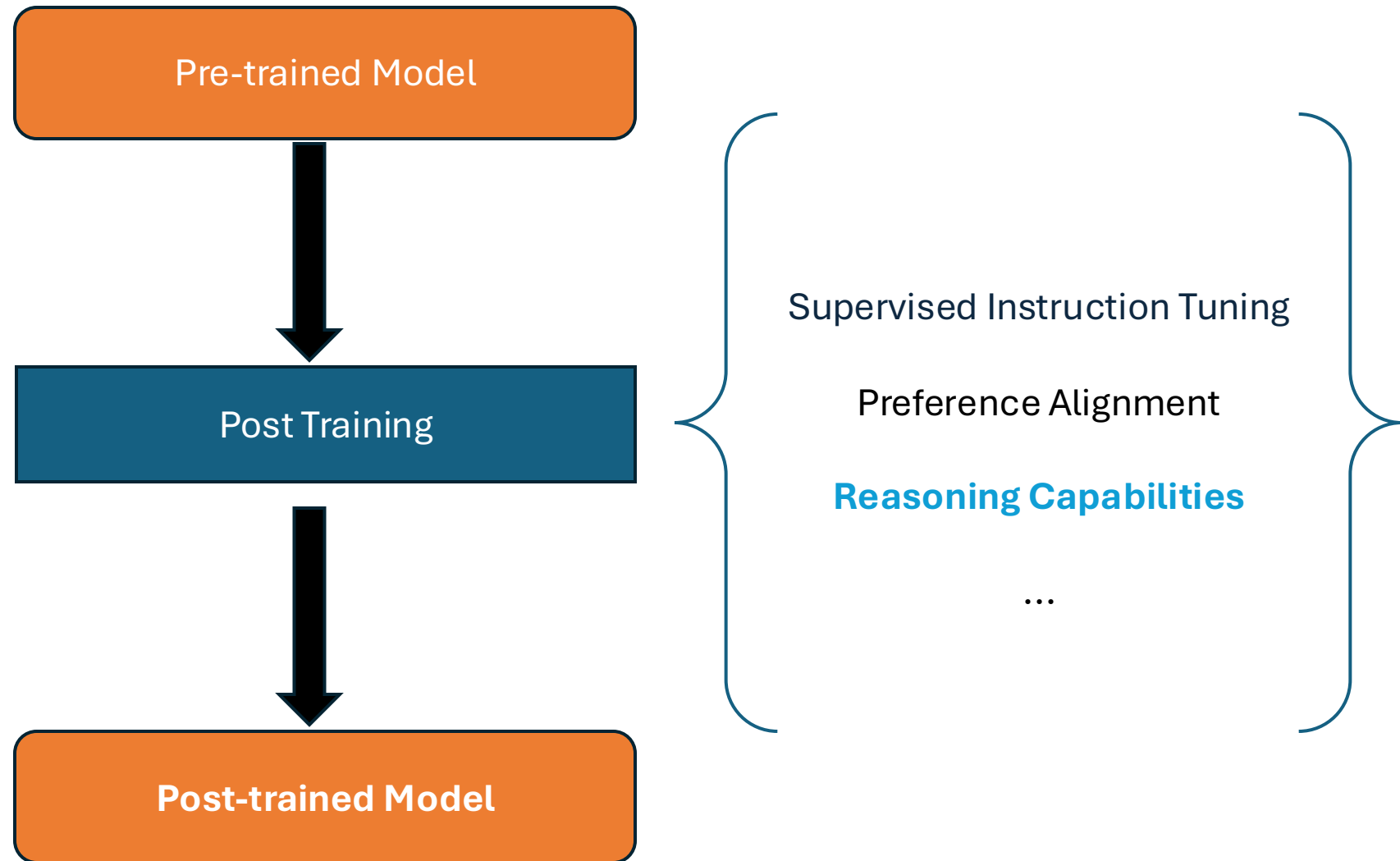


Figure 4: Consistency (Cohen's Kappa) of LLMs' judge results between English and other languages across four datasets and two judge criteria, *Yes / No* and *Grade*.

Summary

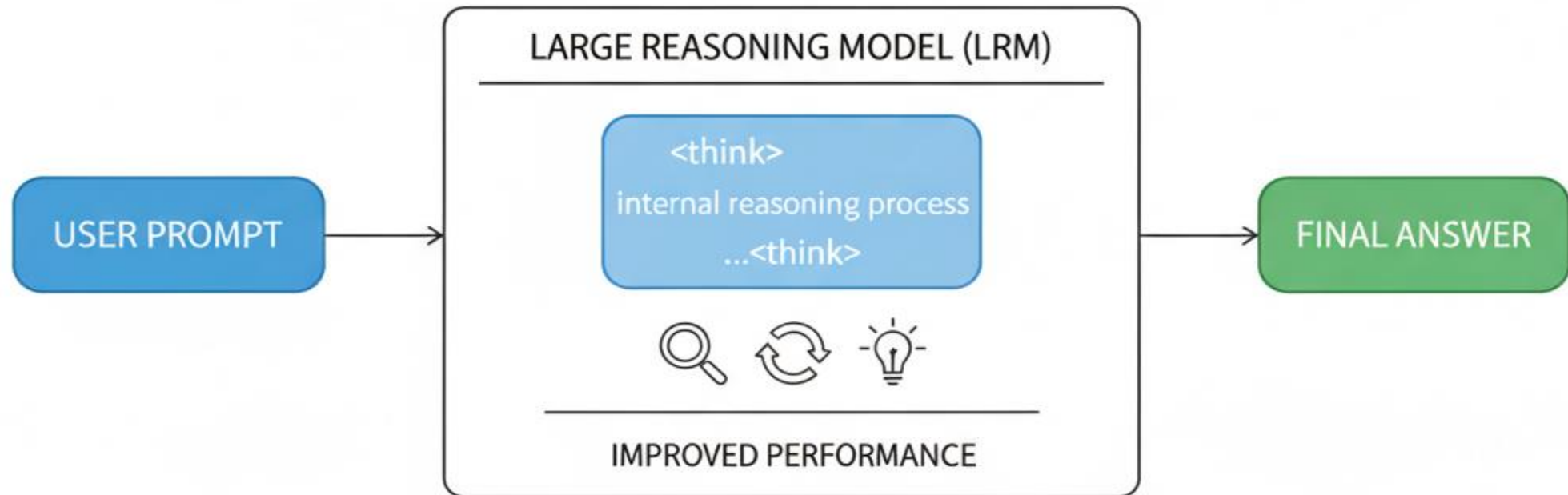
- More languages during RL helps generalize to unseen languages
- In absence of target language Reward model, English reward model helps
- More instances improves alignment
- Have cross-lingual data during RL for better alignment

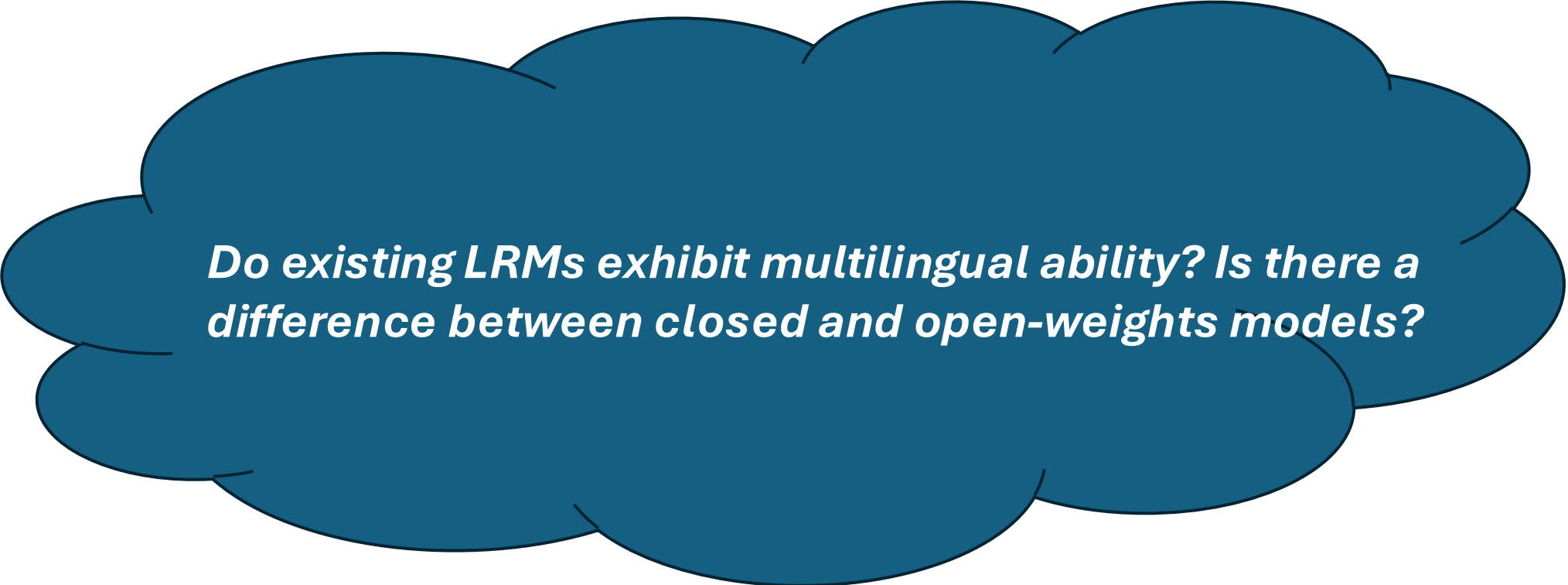


Multilingual Reasoning

What is Reasoning in LLMs?

- A specialized LLM trained to explicitly perform multi-step reasoning (aka LRM)
- Modern Large Reasoning Models (LRMs) have an explicit "thinking" phase, often separated by special tokens like `<think>...</think>`.





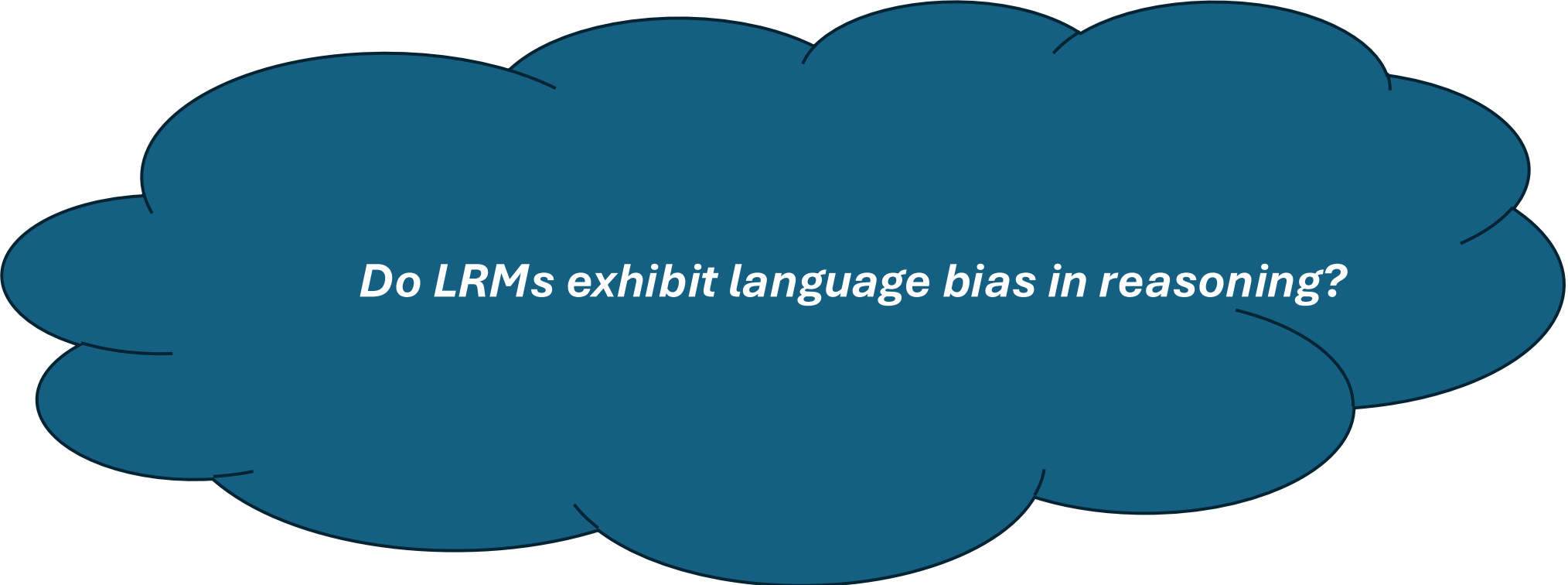
Do existing LRMs exhibit multilingual ability? Is there a difference between closed and open-weights models?

Can LRMs exhibit multilingual capability?

Models	En	De	Es	Fr	Sw	Zh	Bn	Ja	Ru	Th	Avg.
Close-Source LLMs											
ChatGPT-Zero shot	52.0	46.8	52.0	45.6	30.0	44.4	4.4	38.8	37.2	8.0	35.9
ChatGPT-En 2shot	67.2	62.0	61.2	59.2	40.0	52.8	7.6	46.8	50.4	15.6	46.3
GPT4-En 2shot	80.0	73.6	71.2	72.0	64.4	70.0	17.6	71.6	64.0	40.4	62.5
Open-Source LLMs (7B Model)											
LLaMA 2	43.2	37.2	32.4	34.4	5.2	22.4	3.2	15.2	28.0	4.8	22.6
RFT	44.8	33.6	34.0	34.0	2.8	16.8	2.4	6.8	29.2	2.0	20.6
Open-Source LLMs (13B Model)											
LLaMA 2	50.4	42.8	45.2	40.8	7.6	32.8	6.0	25.2	39.2	6.8	29.7
RFT	52.0	38.4	46.8	44.8	3.6	33.6	3.2	26.4	41.6	4.4	29.5
MAmmoth	56.4	45.6	50.0	39.6	1.6	31.2	3.6	19.2	36.8	5.2	28.9
WizardMATH	52.8	40.4	45.6	42.0	5.6	28.0	6.4	22.0	34.4	5.6	28.3

Closed-source models are better at handling non-English CoT

[xMR]

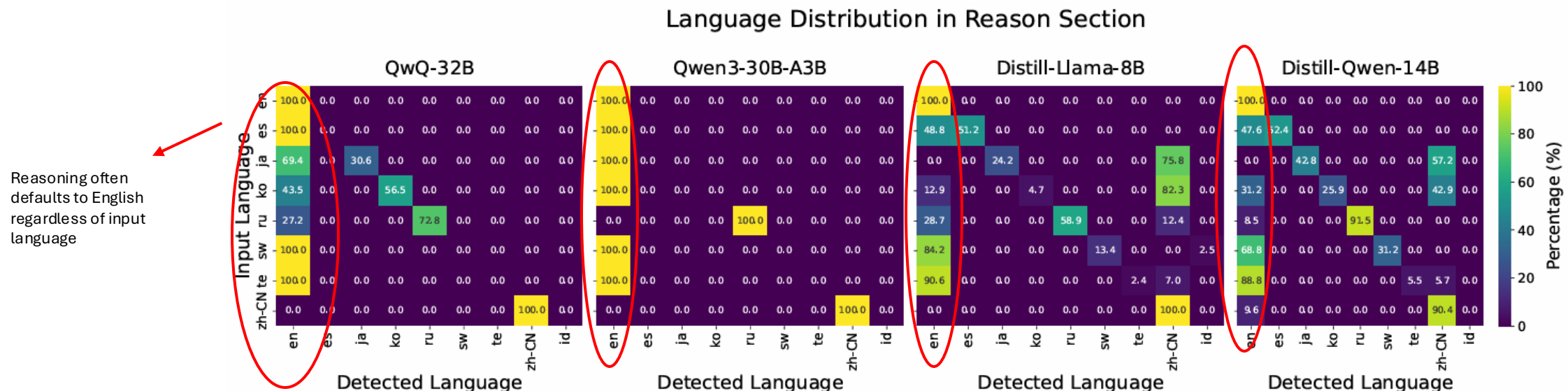


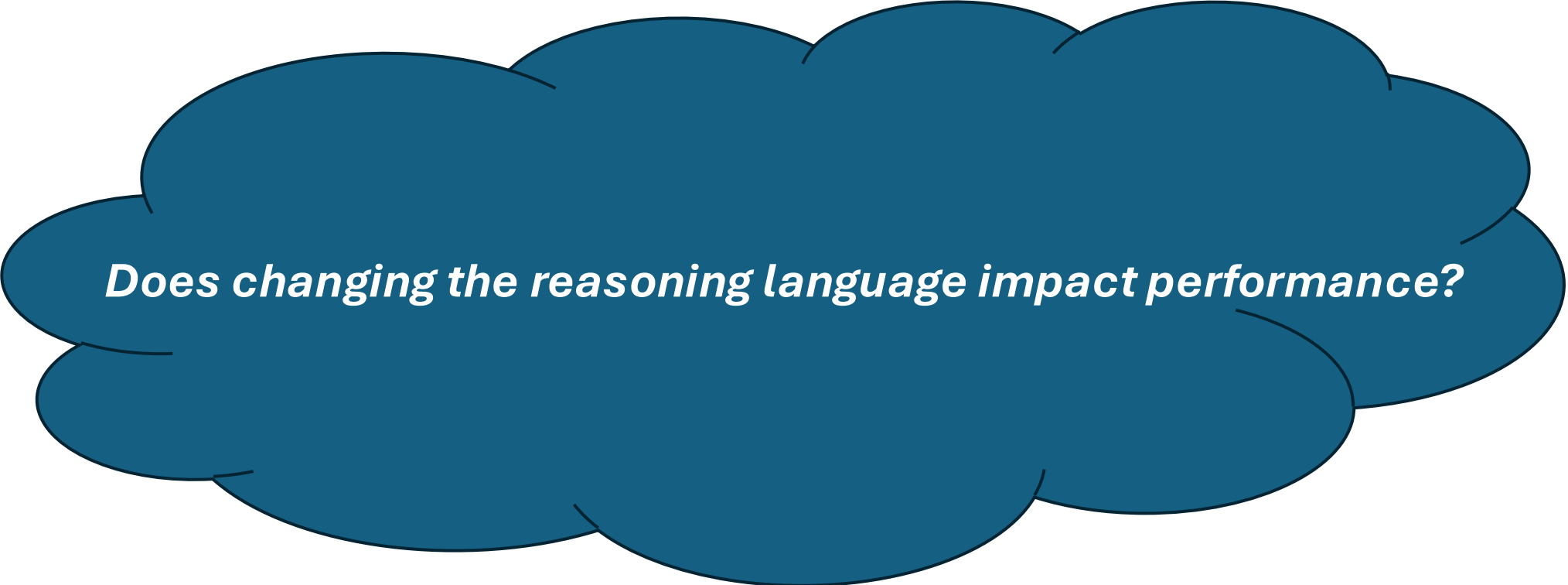
Do LRMs exhibit language bias in reasoning?

Do Models think in English?

[xReasoning]

- Most open-weight models often default to a "reasoning hub" language - primarily **English**, regardless of the input language.
- Closed-source models reason better in non-English languages compared to open-weight models **but there is still a gap**

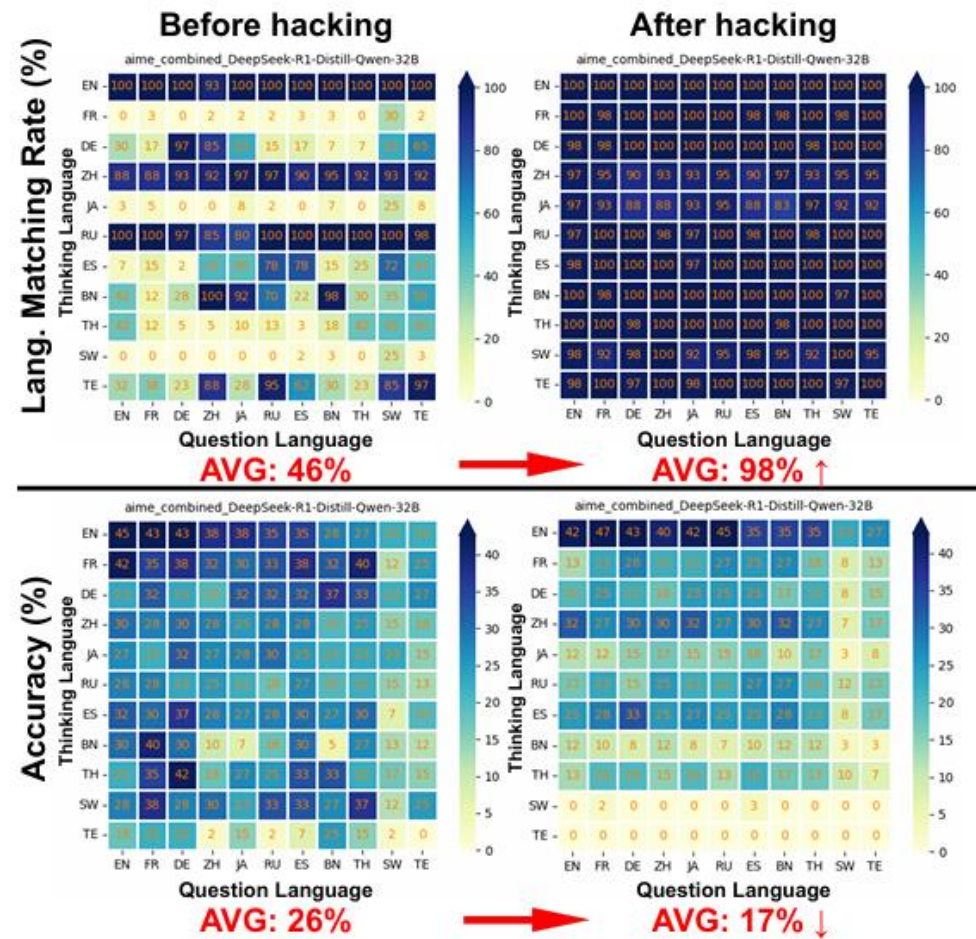




Does changing the reasoning language impact performance?

Effect of Forcing to Think in Target Language

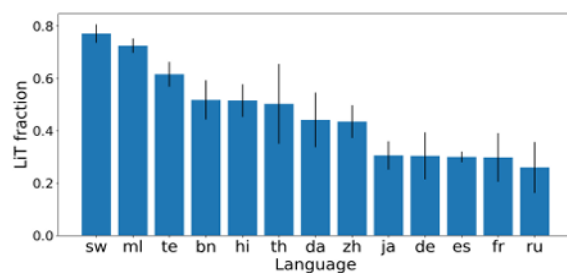
- Forcing non-English reasoning often hurts performance.
- Can we not let it use English?
Reasoning in the prompt language helps
 - Avoid Cross-lingual semantic drift (translation)
 - Language-specific concepts
 - more reliability and
 - trust to the end-user; hence important.



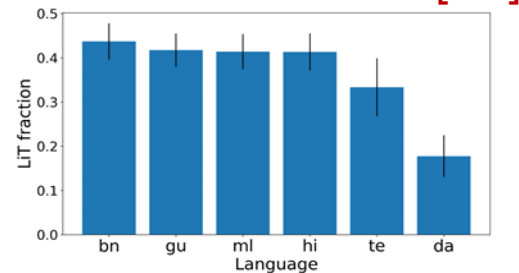
[xReasoning]

Is English Thinking always better?

- English CoT is often better; but not always
- Cells marked in red perform better compared to English when prompt+reasoning is in native language
- **Lost in Translation:** When English CoT is wrong, it is often because the LLM has errors in implicitly translating the prompt from source language



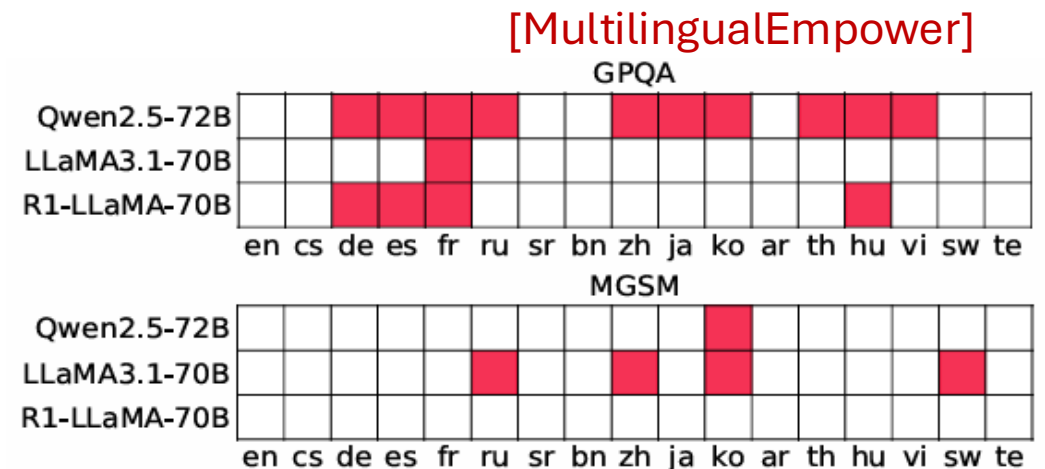
(a) MGSM



(b) GPQA

[LiT]

Figure 5: **Lost in Translation.** Visualization of fraction of incorrect answers occurring due to errors in translation when reasoning in English. Error bars represents standard deviation.



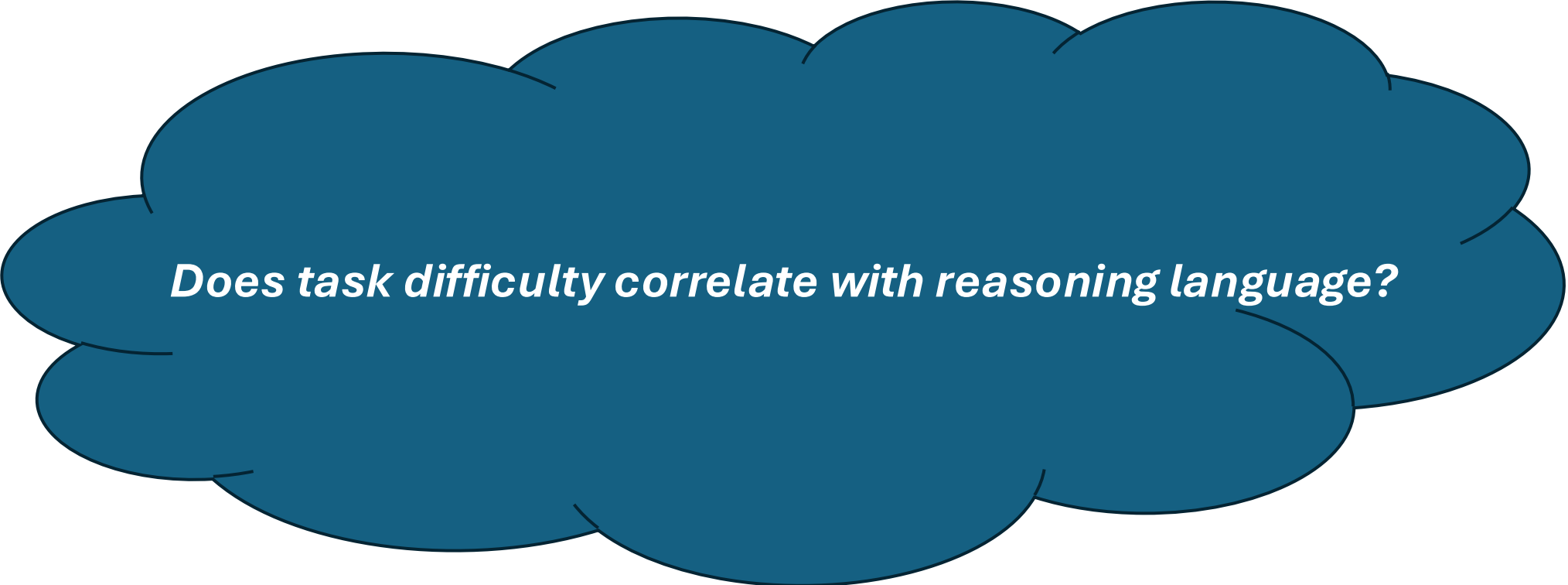
Is English Thinking always better?

- Forcing reasoning in the **native language** can **improve performance** on cultural knowledge benchmarks and **reduce toxicity** in safety evaluations.

[LanguageMatters]

Strategy	English	Chinese	Spanish	Swahili	Japanese	Korean
Prefill English (EN)	–	83.1%	83.8%	48.8%	80.8%	77.6%
Prefill Target Language	83.0%	80.2%	78.7%	35.3%	74.0%	71.2%
Difference (EN - Target)	–	+2.9%	+5.1%	+13.6%	+6.8%	+6.4%

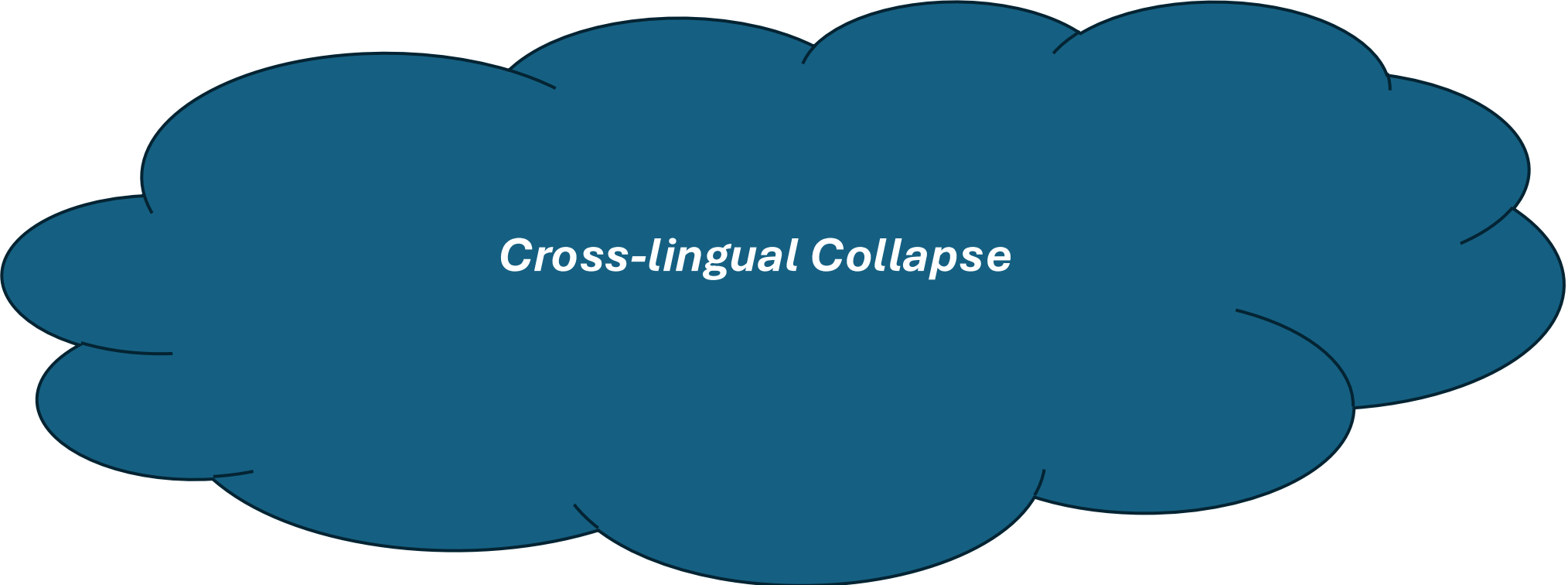
Strategy	Chinese	Spanish	Russian	Swahili	Japanese	Telugu	Korean
Average across all models							
Prefill English (EN)	7.7%	12.3%	11.5%	3.5%	9.9%	0.8%	4.6%
Prefill Target Language	7.4%	13.3%	16.1%	3.6%	9.5%	1.6%	3.8%
Difference (EN - Target)	+0.3%	-1.0%	-4.6%	-0.1%	+0.4%	-0.8%	+0.8%



Does task difficulty correlate with reasoning language?

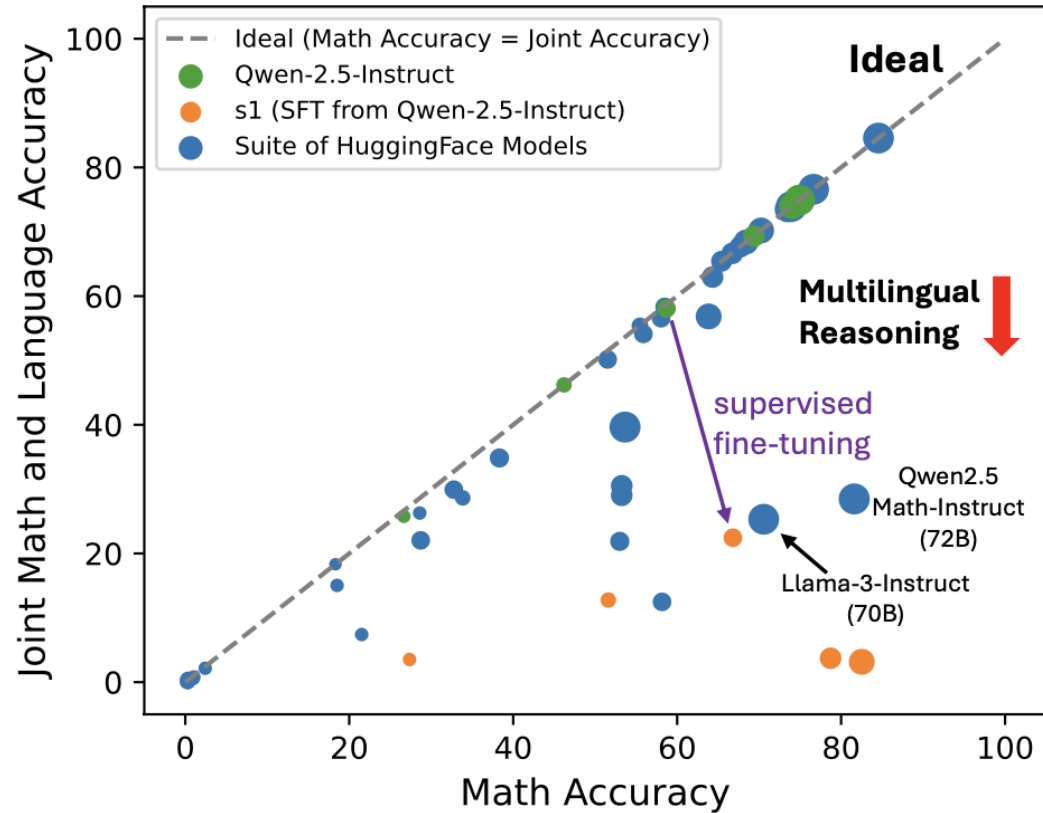
Impact of Task Difficulty on Reasoning Language

- For easy tasks (e.g., MATH500), language choice has little effect on accuracy, while non-English reasoning uses fewer tokens
- As task difficulty increases, we see a relative drop of upto 50% on average compared to English
- For harder tasks, high-resource languages (Zh, Ru), outperform low-resource one (Hi).



Cross-lingual Collapse

Cross-lingual Collapse



- A Multilingual model reverts to reasoning in dominant language even though the prompt is in different language
- The LLM prefers English reasoning paths as it maximizes the rewards

[CrossLingualCollapse, GapsInMR]

M2A: Overcoming Multilingual Collapse

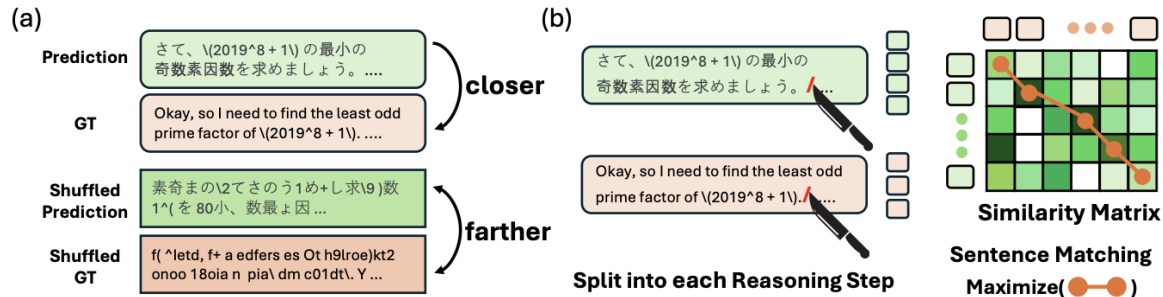


Figure 3: **Overview of M2A.** (a) *Multilingual Context Alignment* enforces global similarity between generated and reference responses while discouraging trivial matches via shuffled negatives. (b) *Multilingual Reasoning-Step Alignment* provides finer-grained supervision by aligning individual reasoning steps with ground-truth traces using dynamic programming.

[M2A]

- Given question (q) and English ground truth (y)
 - Translate q to each target language
 - Generate multiple outputs for each translated question
 - Calculate reward for each output
- Multilingual Context Alignment
 - Use cosine distance between y and generated output as rewards
 - Shuffle output and ground truth to get negative rewards [ensure language is different]
- Multilingual Reasoning Step Alignment
 - Split y and output into sentences
 - Align and calculate rewards at the sentence level
- Language Consistency Reward

M2A: Overcoming Multilingual Collapse

Table 2: Accuracy of Qwen2.5-7B-Instruct and post-training methods in GSM8K (English) and MGSM (ten languages). Results are reported for mathematical accuracy (Math.), language accuracy (Lang.), and joint accuracy (Joint). Bold indicates the best performance in each column.

Method	GSM8K			MGSM		
	Math.	Lang.	Joint	Math.	Lang.	Joint
Qwen-2.5-Instruct	81.2	100	81.2	58.7	99.0	58.1
GRPO	80.4 $\pm$ 0.9	100.0 $\pm$ 0.0	80.4 $\pm$ 0.9	58.8 $\pm$ 0.4	95.9 $\pm$ 2.9	58.2 $\pm$ 0.7
SFT (s1)	87.2 $\pm$ 1.6	100.0 $\pm$ 0.0	87.2 $\pm$ 1.6	66.7 $\pm$ 0.1	31.0 $\pm$ 0.5	21.9 $\pm$ 0.6
SFT on s1K-X	84.3 $\pm$ 1.1	66.7 $\pm$ 33.3	56.5 $\pm$ 28.3	45.2 $\pm$ 4.1	99.7 $\pm$ 0.1	45.0 $\pm$ 4.3
M2A (ours)	87.3 $\pm$ 0.1	100.0 $\pm$ 0.0	87.3 $\pm$ 0.1	59.0 $\pm$ 0.3	97.8 $\pm$ 0.2	58.1 $\pm$ 0.4

[M2A, GapsInMR]

Table 5: Contribution of individual reward functions to M2A. The evaluation is performed on GSM8K and MGSM. Bold means the best performance. Lang: Language Consistency, CA: Context Alignment, RA: Reasoning-Step Alignment.

M2A Variants			GSM8K			MGSM		
Lang	CA	RA	Math.	Lang.	Joint	Math.	Lang.	Joint
✓			86.9 $\pm$ 0.0	100.0 $\pm$ 0.0	86.9 $\pm$ 0.0	54.2 $\pm$ 0.1	98.3 $\pm$ 0.1	53.8 $\pm$ 0.1
✓	✓		84.7 $\pm$ 0.1	100.0 $\pm$ 0.0	84.7 $\pm$ 0.1	57.8 $\pm$ 0.1	99.5 $\pm$ 0.1	57.5 $\pm$ 0.1
✓	✓	✓	87.3 $\pm$ 0.1	100.0 $\pm$ 0.0	87.3 $\pm$ 0.1	59.0 $\pm$ 0.3	97.8 $\pm$ 0.2	58.1 $\pm$ 0.4

- SFT
 - improves accuracy in English and non-English languages
 - Reasoning is generated in English language
- SFT on Translated Data
 - drop in non-English language accuracy
 - Improves response generation in non-English language
- M2A
 - Improves all metrics
 - Accuracy in non-English language is lower compared to SFT

M-Thinker: Overcoming Cross-lingual Collapse

Metric: Language Consistency & Accuracy (LC&Acc, %)													
DeepSeek-R1-Distill-Qwen-1.5B	0.22	0.02	7.05	11.92	0.12	3.87	46.56	13.38	0.16	3.56	32.30	19.19	11.53
Prompt-Control (No Training)	0.98	0.02	9.69	17.34	0.22	5.65	46.42	19.65	0.62	13.52	31.40	22.32	13.99
DIT (No Training)	0.32	0.00	4.18	14.58	0.12	3.84	45.93	11.69	0.40	4.82	33.76	19.32	11.58
QRT (No Training)	4.94	0.14	13.10	23.52	0.52	8.45	46.22	23.75	1.39	13.91	35.42	24.14	16.29
Cold-Start SFT	1.11	0.00	17.29	16.99	1.56	7.39	45.84	20.54	0.52	7.25	34.51	21.73	14.56
Naive-RL	0.00	0.00	0.00	0.00	0.00	0.00	55.31	0.00	0.00	0.00	25.47	16.16	8.08
SLC-RL	0.00	0.00	0.00	0.00	0.00	0.00	56.37	0.00	0.00	0.00	40.99	19.47	9.74
M-Thinker-1.5B $\Rightarrow$ Iter-1 (Ours)	<u>34.25</u>	<u>24.48</u>	<u>43.72</u>	<u>45.78</u>	<u>28.72</u>	<u>35.39</u>	54.89	<u>49.19</u>	6.39	<u>35.76</u>	<u>45.60</u>	<u>38.37</u>	<u>36.88</u>
M-Thinker-1.5B $\Rightarrow$ Iter-2 (Ours)	45.54	32.86	49.75	51.47	32.72	42.47	56.41	49.20	<u>2.80</u>	37.55	48.20	38.83	40.65

Table 1: The LC, Acc, and LC&Acc (%) results on the MMATH benchmark of the DeepSeek-R1-Distill-Qwen-1.5B backbone. “*ID-avg/OOD-avg*” is the average result of five In-Domain/Out-of-Domain languages and “*All-avg*” is the average result of all ten languages. The result in **bold** means the best result, and the underlined result means the second-best result in each setting. “**Iter-1/2**” means the training iteration 1/2.

Rewards

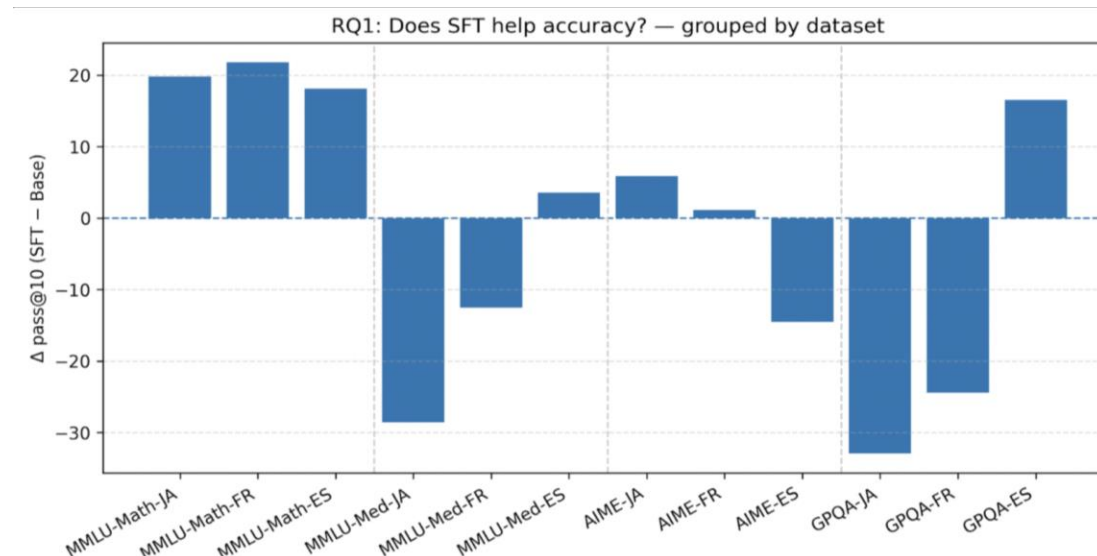
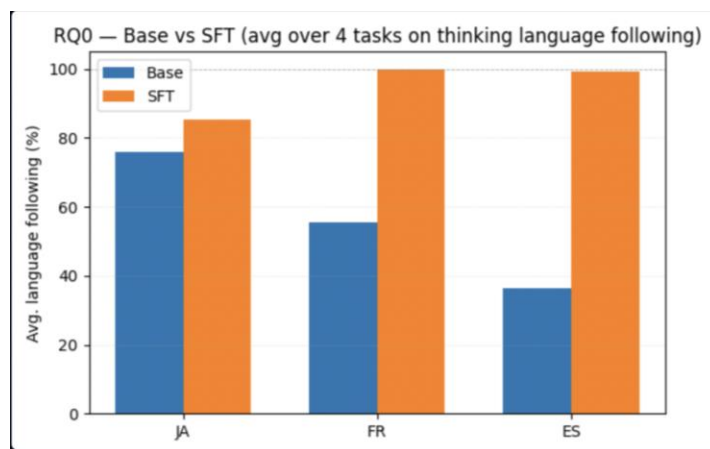
- Language Consistency Reward
- Cross-lingual Thinking Alignment Reward
 - Use a LLM to score English and target language reasoning alignment
- Format Reward
- Accuracy Reward

Reward hacking helps avoid cross-lingual collapse to a certain extent

[M-Thinker]

Budget Alignment: Making LLMs reason in Your Language

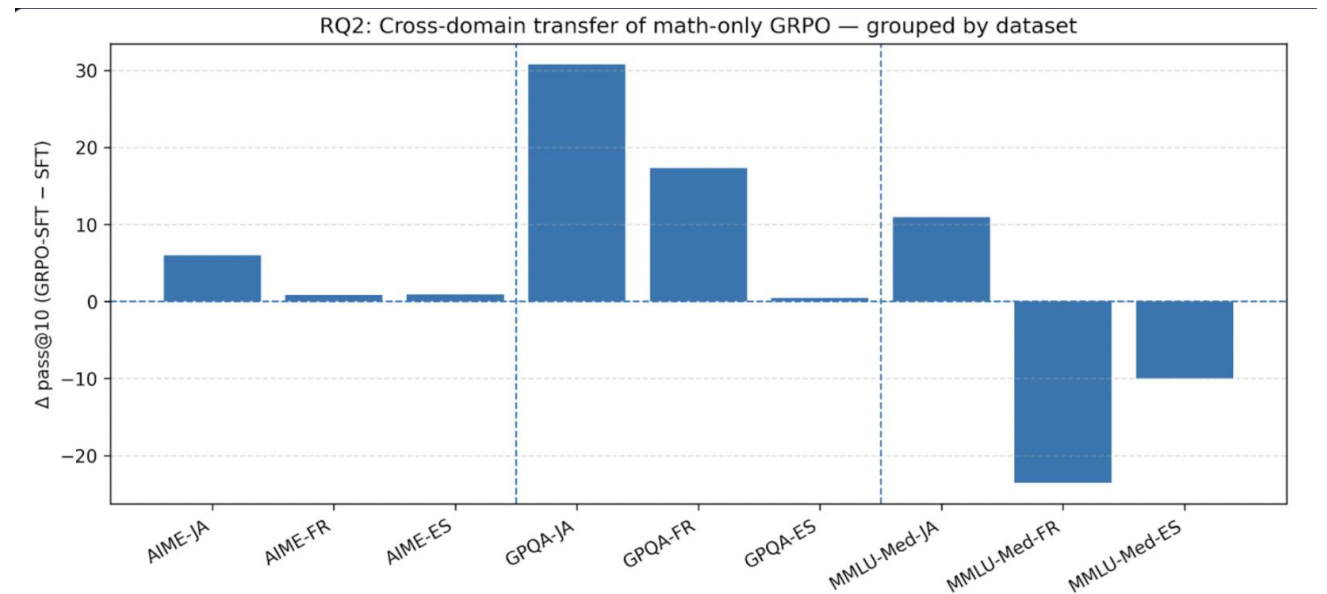
1. Perform SFT using small high-quality in-language data
2. In-domain improves --> out of domain drop



[BudgetAlign]

Budget Alignment: Making LLMs reason in Your Language

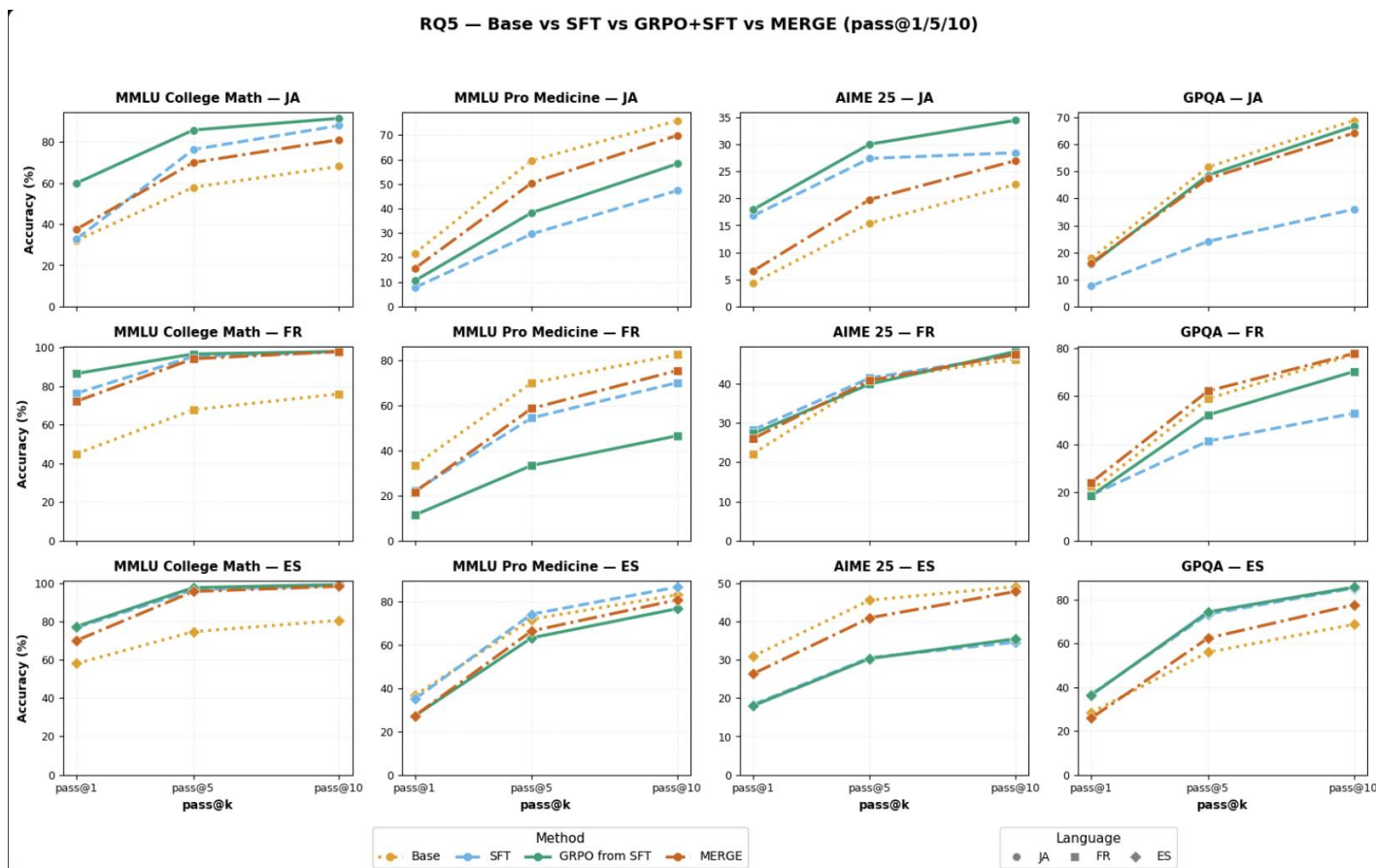
1. Do GRPO-SFT on the stage 1 model
 - Use high-clip and no KL – ensures the model doesn't revert to English for reasoning
2. GRPO improves on SFT while retaining language accuracy (out-of-domain too)



[BudgetAlign]

Budget Alignment: Making LLMs reason in Your Language

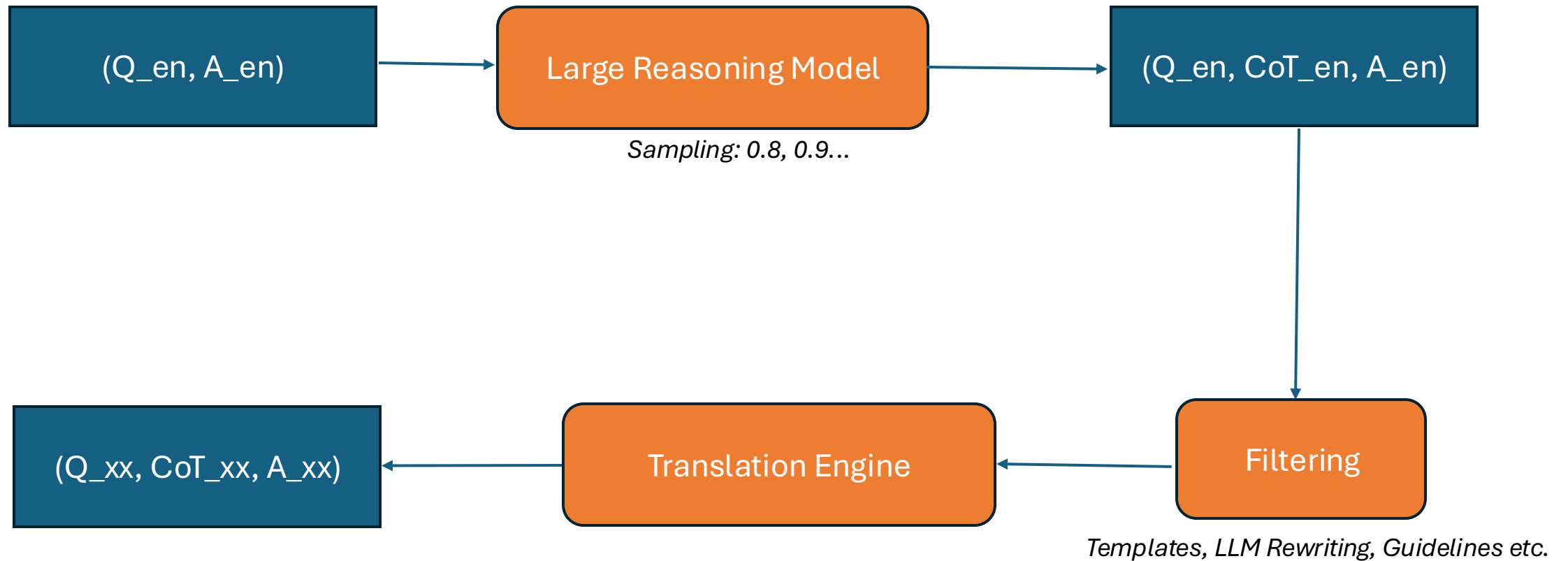
1. Merging helps improve performance on other tasks
2. Slight drop-in performance on in-domain



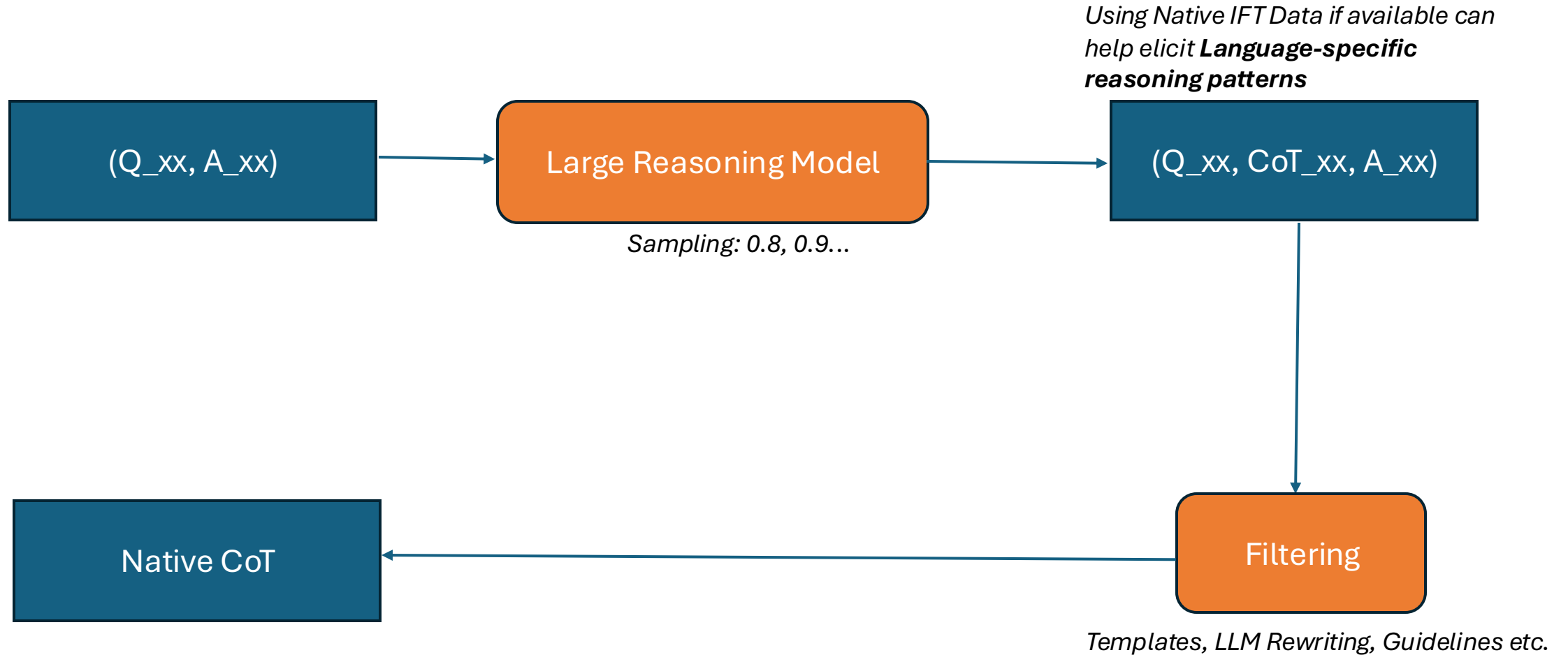
[BudgetAlign]

Training Multilingual Reasoning Models by Imitating

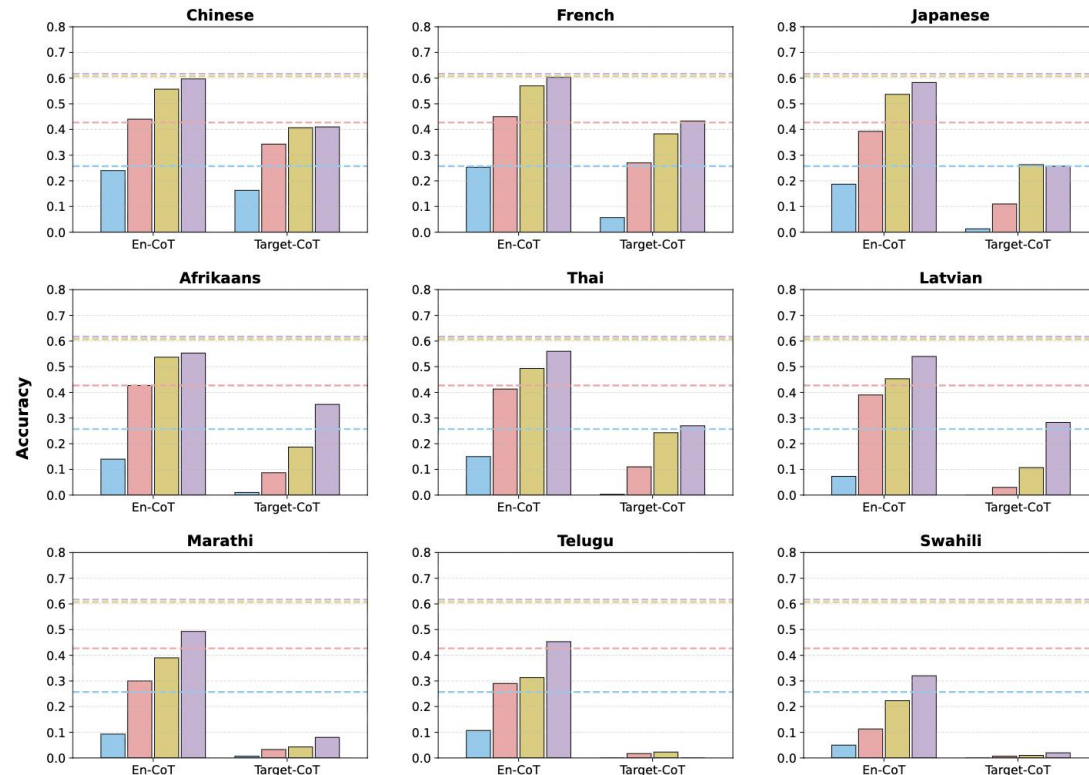
Distilling Native CoT From English



Distilling Native CoT Data Directly



Translated English CoT vs Target-language CoT



[CoTXLanguages]

- High-resource languages
 - Scaling model improves reasoning in both English and Target language CoT
- Low-Resource Languages
 - Translated English CoT is better than target language CoT
 - Target Language CoT with larger models fails to reach English performance with smaller models

Finding: Translating English CoT to target language and then finetuning is better than generating CoT in native language (**this is task-dependent**)

Summary and Recommendations

- LRMs can reason better in English irrespective of the query language
- For culturally-sensitive queries, query language reasoning is better than English
- Enabling LRMS to reason in query language is an active area of research
 - SFT followed by GRPO leads to better reasoning in target language
- Finetuning using translated reasoning traces is better than finetuning on native reasoning traces

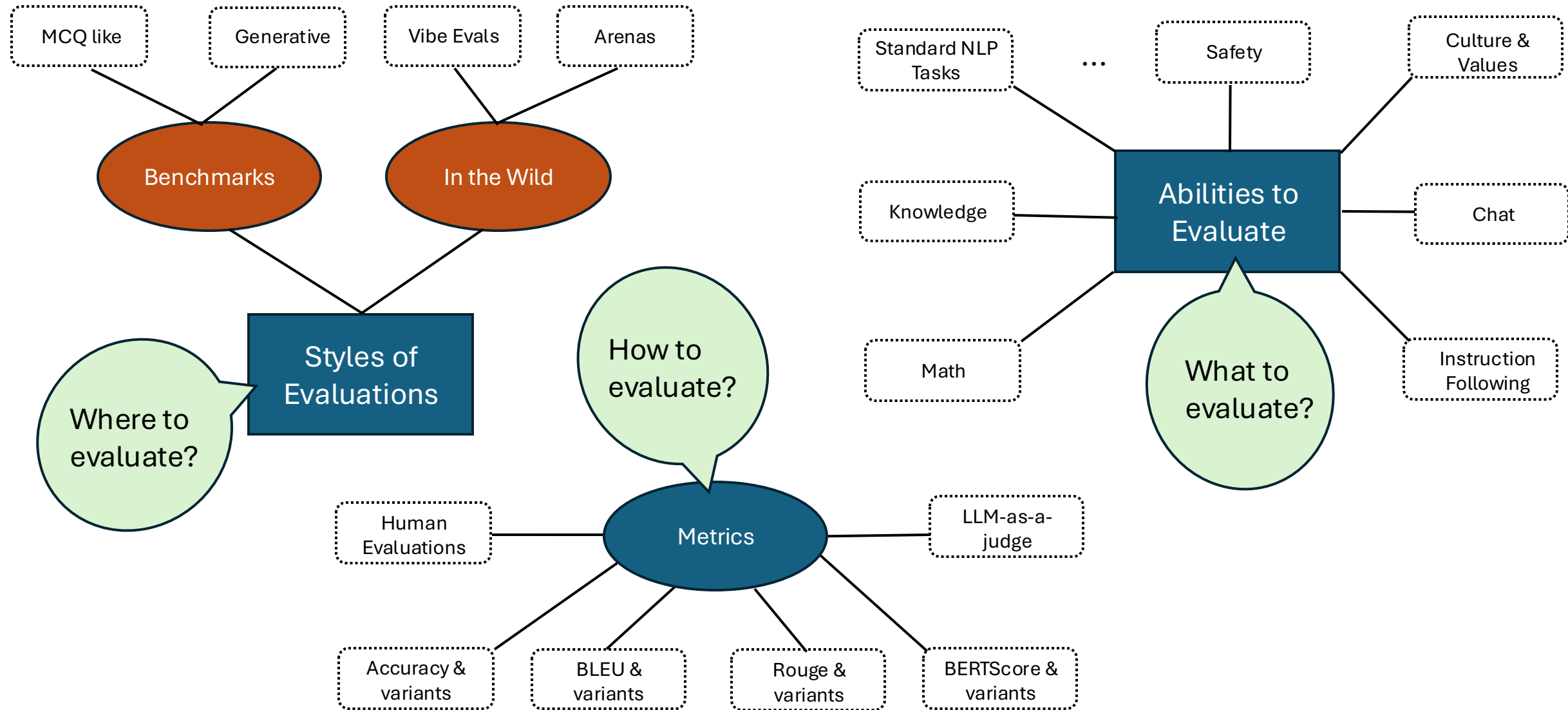
Tutorial Outline

- Introduction
 - The Multilingual LLM Challenge
- Extending LLMs to new languages
 - Vocabulary Expansion
 - Continued Pre-training
 - Post-Training
- **Multilingual Evaluation**
- Summary & Future Directions

Multilingual Evaluations

How to measure our model improvements?

Broad Taxonomy of Evaluations

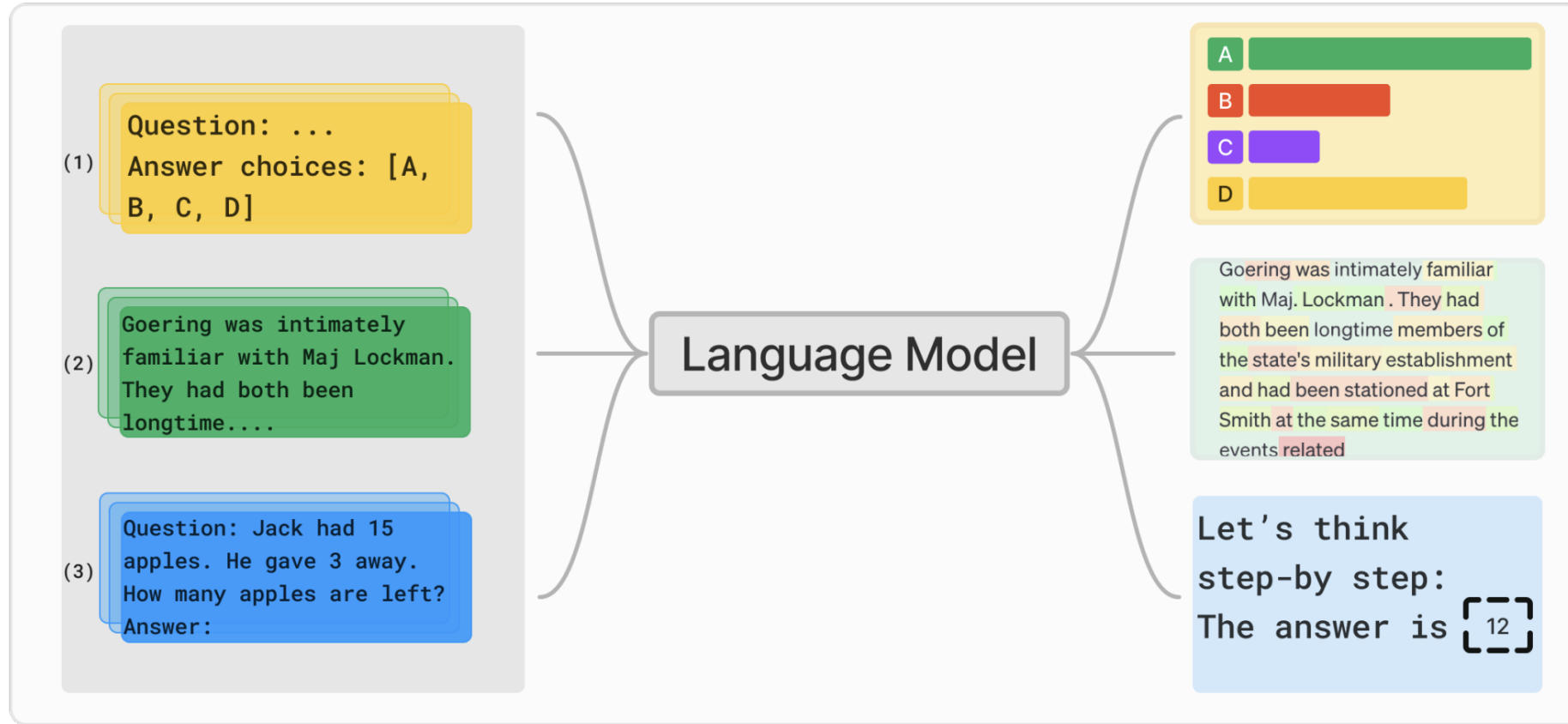


Modern LM Evals - Benchmarks

Overview of the Capabilities to Measure

Name	Description	Famous Benchmarks
Knowledge, Academic	Measures the model's grasp of facts and diverse academic subjects.	MMLU, MMLU-Pro,
Commonsense Reasoning	Tests the model's understanding of unstated, everyday world knowledge.	Hellaswag, Winogrande, BoolQ, PiQA, CommonsenseQA, etc
Reading Comprehension	Assesses the ability to read a text passage and answer questions about it.	SQuAD, RACE
Math	Measures proficiency in solving mathematical problems (word problems, etc.).	GSM8K, MATH, MATH-500
Reasoning	Probes logical deduction, problem-solving, and multi-step thinking.	ARC, DROP, GPQA, LogiQA
Coding	Evaluates the ability to write, debug, and understand computer programs.	HumanEval, MBPP
Instruction Following	Tests if the model can accurately follow specific and complex user commands.	AlpacaEval, IFEval, Arena-Hard
Conversation Ability	Assesses the quality of multi-turn dialogue, coherence, and helpfulness.	MT-Bench, WildBench, BigGenBench
Safety, Fairness, Robustness	Measures adherence to safety guidelines and resilience to bias/adversarial inputs.	BBQ, ToxiGen, TruthfulQA, XS-Test, Adv-Bench
Tool/Function Calling	Evaluates the ability to correctly use external tools or APIs to complete a task.	ToolBench, MCP-Bench, API-Bank
Domain/Application Specific	Tests expertise in a specialized field, such as medicine or law.	MedQA, LegalBench,

Evaluation Paradigms



**Conditional
Log Likelihood**

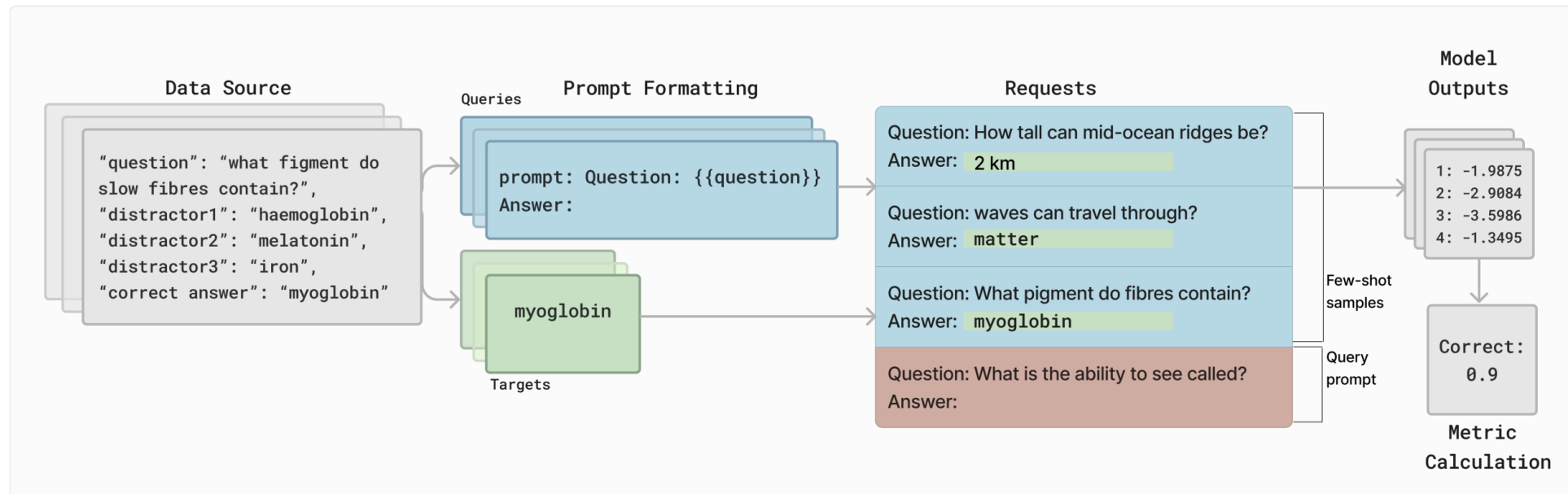
**Rolling
Log Likelihood (Pplx)**

Generate (until)

Frameworks for Evaluation

- Lm-evaluation-harness

Solves the
orchestration problem



Language Model Evaluation Harness

[LMEvalHarness]

DOI [10.5281/zenodo.10256836](https://doi.org/10.5281/zenodo.10256836)

Frameworks for Evaluation

- Evalchemy
- OpenAI Evals
- DeepEval
- Opik
- LightEval
- Etc...

 Evalchemy

A unified and easy-to-use toolkit for evaluating post-trained language models

DeepEval.

The LLM Evaluation Framework

 **opik**

Opik

Open-source LLM evaluation platform

 **Lighteval**

Your go-to toolkit for lightning-fast, flexible LLM evaluation, from Hugging Face's Leaderboard and Evals Team.

Tests passing Quality passing python 3.10 License MIT pypi v0.11.0

Now let's see the Multilingual
World

Current State of Multilingual Benchmarks

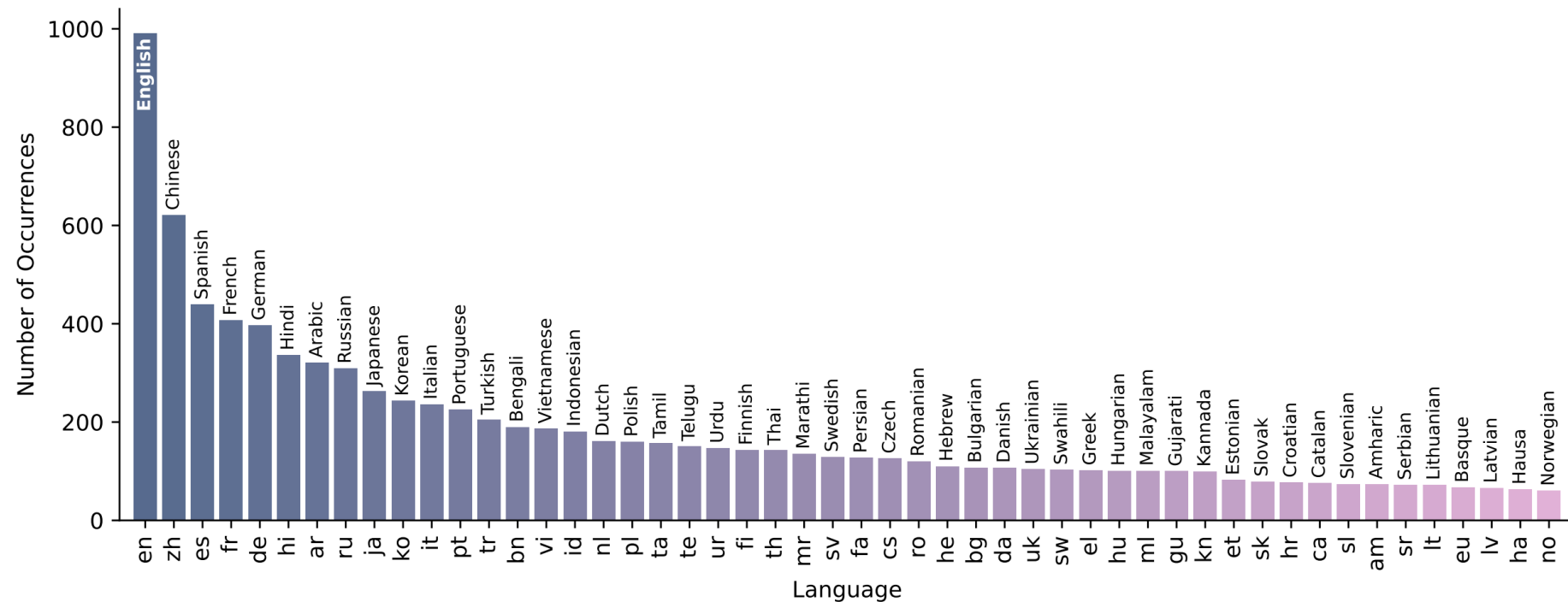
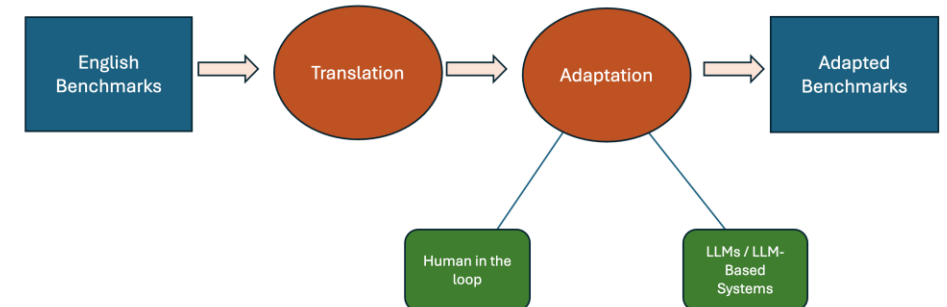
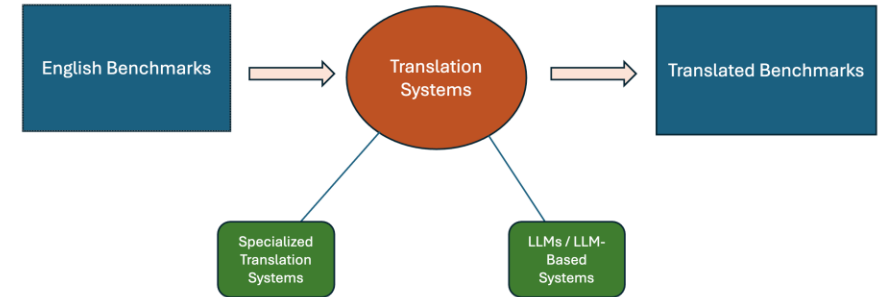


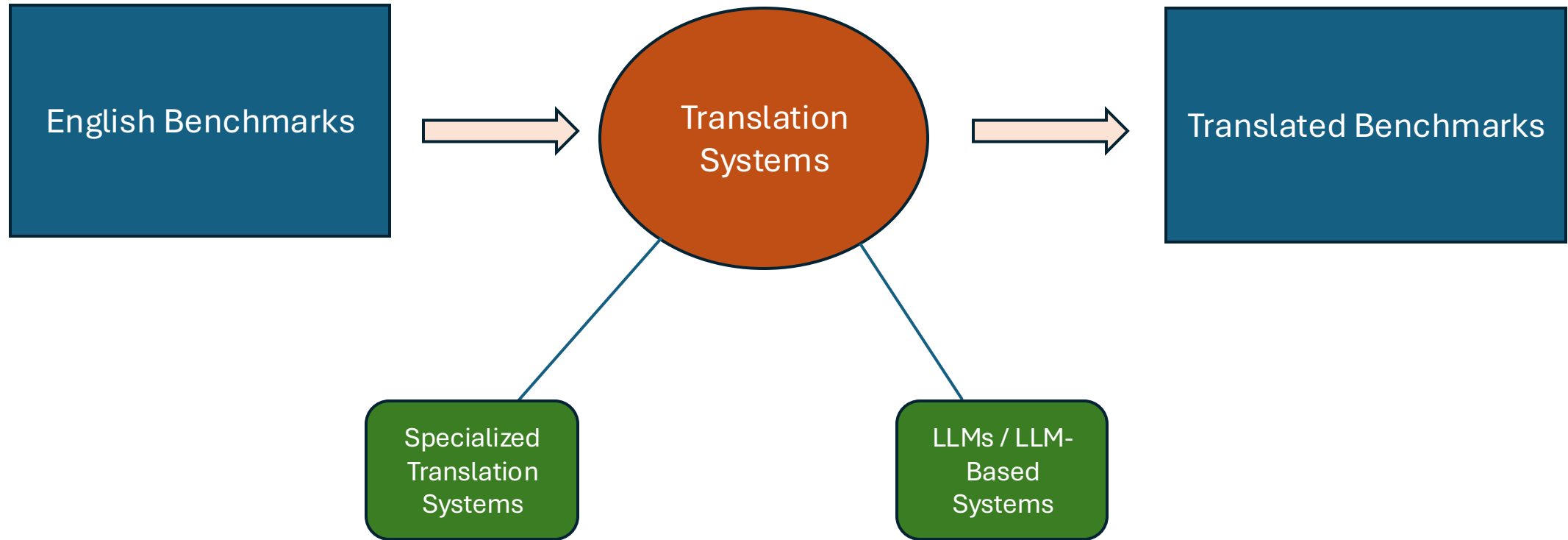
Figure 2 | Distribution of the top 50 languages in our multilingual benchmark collection. Although English is deliberately excluded from the collection, it still appears as the most frequent language in the collection. This distribution illustrates the current imbalance in multilingual evaluation benchmarks.

Strategies for Creating Benchmarks

- Machine Translation
- Adaptation
- Create from Scratch



Translation



Translation (Benefits)

1. Quick and Easy way
2. Easily leverage existing high quality test-sets
3. Enables fair cross-lingual comparison.

Translation (Challenges)

1. **Translationese:** Introduces Semantic Drift, Mistranslation of Culturally-specific content, idioms, etc.

Its raining cats and dogs

यह बिल्लियों और कुत्तों की
बारिश हो रही है।

मूसलाधार बारिश हो रही है।

Translation (Challenges)

- **Loss of Cultural Grounding:** Translation inherits English-centric knowledge and worldview, failing to reflect real-world relevance.

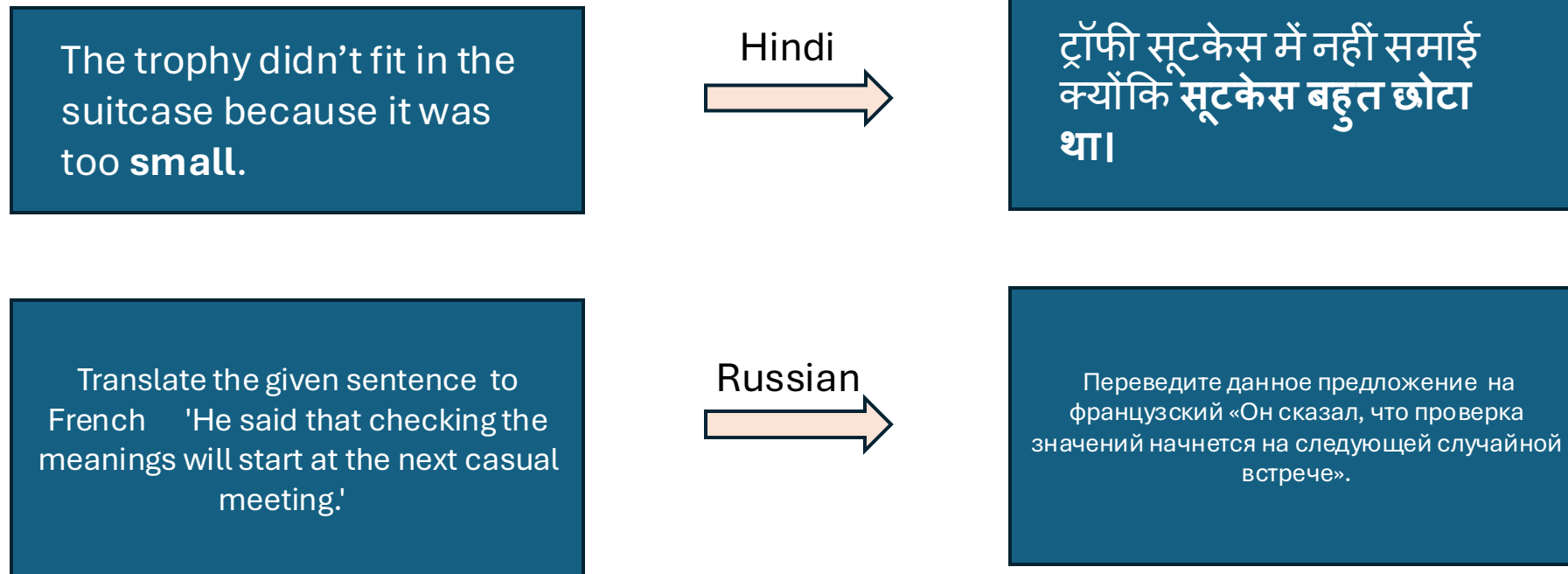
Which of the following
best describes a
contributing factor in the
crafting of the United
States Constitution?

VS

Which of the following
best describes the role of
the Drafting Committee,
chaired by Dr. B. R.
Ambedkar, in the
formation of the Indian
Constitution?

Translation (Challenges)

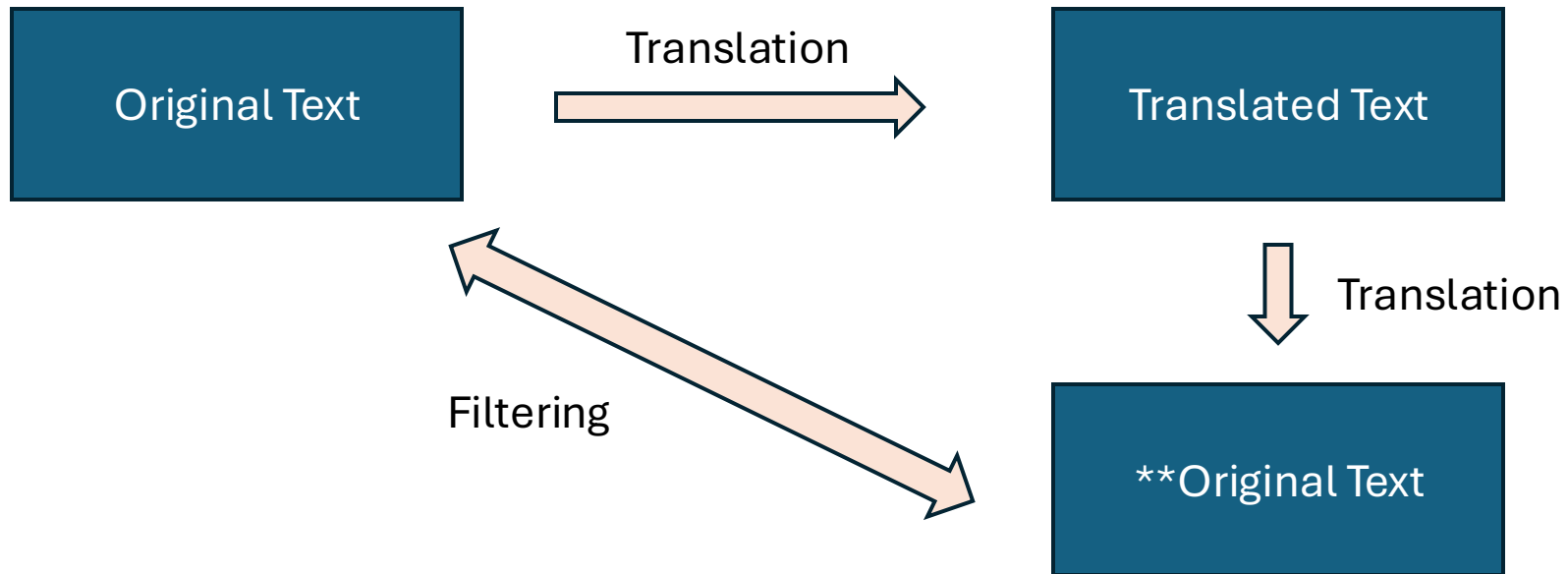
- **Difficulty Preserving Task Properties:** Some tasks lose their discriminative properties upon translation – making them easier or unintended ways?



Translation (Challenges - Solutions)

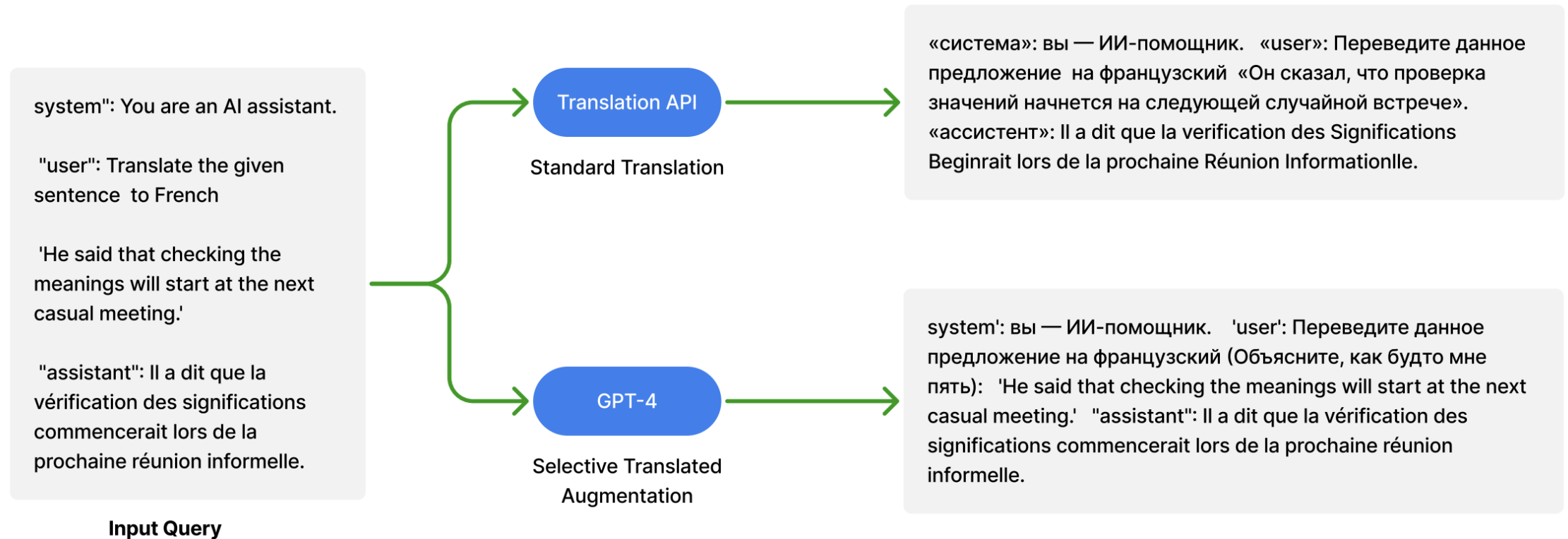
Solution?

- Human-in-the-loop (Expensive but Gold)
- Backtranslate



Translation (Challenges - Solutions)

- LLM-Based Selective/Adaptive Translation



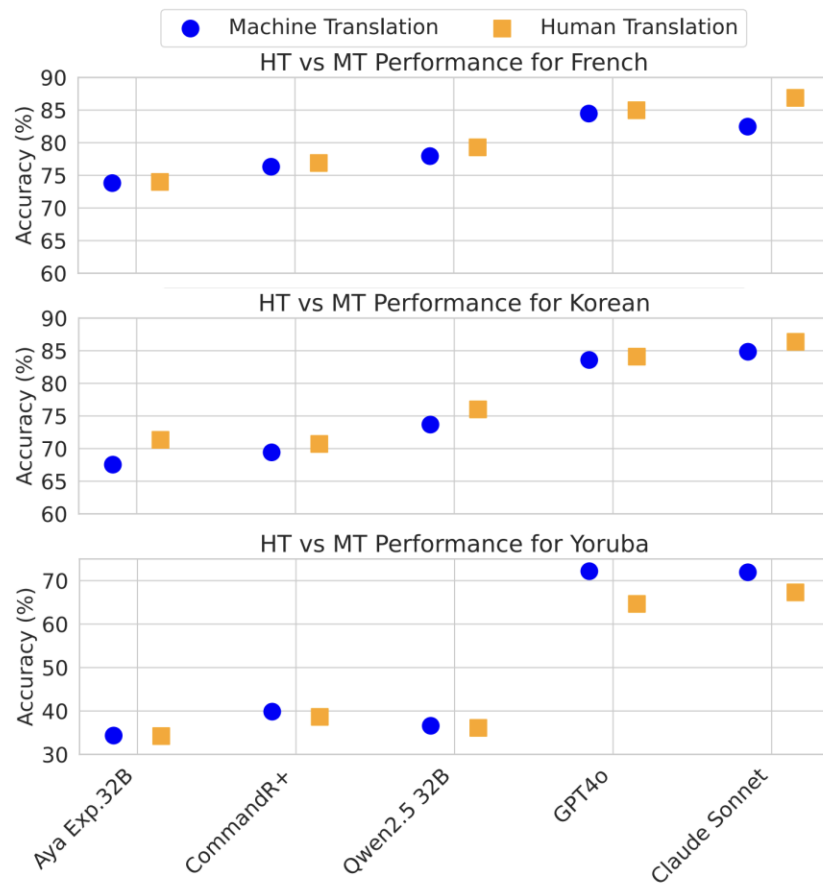
Translation – Still a reasonably good alternative

	ar	bg	de	el	es	et	eu	fr	hi	ht	id	it	ja	ko	my	qu	ru	sw	ta	te	th	tr	ur	vi	zh
XLM-R																									
XCOPA	-	-	-	-	-	69/73	-	-	-	50/57	76/69	75/73	-	-	-	49/59	-	66/58	64/68	-	69/66	71/66	-	68/70	72/73
XStoryCloze	80/80	-	-	-	84/84	-	79/79	-	78/79	-	88/87	-	-	-	71/70	-	84/85	75/76	-	76/75	-	-	-	-	87/86
XNLI	78/78	82/82	82/82	81/80	83/78	-	-	82/83	75/80	-	-	-	-	-	-	-	79/82	70/74	-	-	76/76	77/79	-	78/81	79/74
PAWS-X	-	-	90/91	-	91/91	-	-	91/90	-	-	-	-	81/84	81/83	-	-	-	-	-	-	-	-	-	-	83/82
BLOOMz																									
XCOPA	-	-	-	-	-	52/52	-	-	-	N/A	78/78	62/65	-	-	-	51/52	-	60/63	75/72	-	N/A	50/50	-	80/77	71/67
XStoryCloze	88/88	-	-	-	91/91	-	84/78	-	85/84	-	91/90	-	-	-	54/52	-	73/73	79/79	-	74/73	-	-	-	-	70/70
B-NLI	71/72	66/68	69/68	65/66	73/74	-	-	72/73	70/72	-	-	-	-	-	-	-	69/70	70/71	-	-	N/A	68/70	-	72/73	74/72
AYA																									
XCOPA	-	-	-	-	-	87/84	-	-	-	82/83	87/87	88/88	-	-	-	56/56	-	79/83	86/83	-	84/82	86/85	-	85/84	86/84
XStoryCloze	95/92	-	-	-	94/94	-	83/75	-	93/91	-	91/87	-	-	-	94/86	-	90/82	93/89	-	93/88	-	-	-	-	95/90
B-NLI	78/79	79/79	78/78	78/78	79/80	-	-	79/80	75/75	-	-	-	-	-	-	-	79/79	74/75	-	-	79/79	79/79	-	76/77	77/77

Table 2: The (%) accuracy of the models on the human translated (original)/our machine translated datasets.

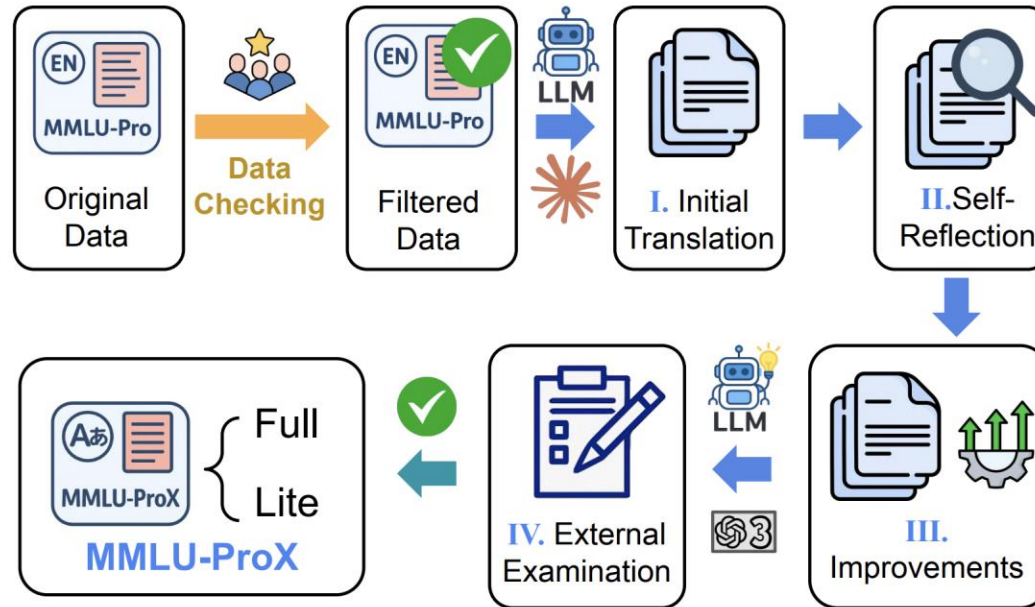
[MTvsHumanEval]

Translation – Still a reasonably good alternative



- Generally perform better on Human Translated data for high resource languages.
- Trend slightly shifts as we move towards low-resource as models primarily rely on machine-translated data for training.

Translation (Existing Approaches)



Completely
automatic LLM
based QC

Figure 2: MMLU-ProX Data Pipeline: A rigorous four-stage process consisting of data curation, translation, external model verification, and expert review.

Translation (Existing Approaches)

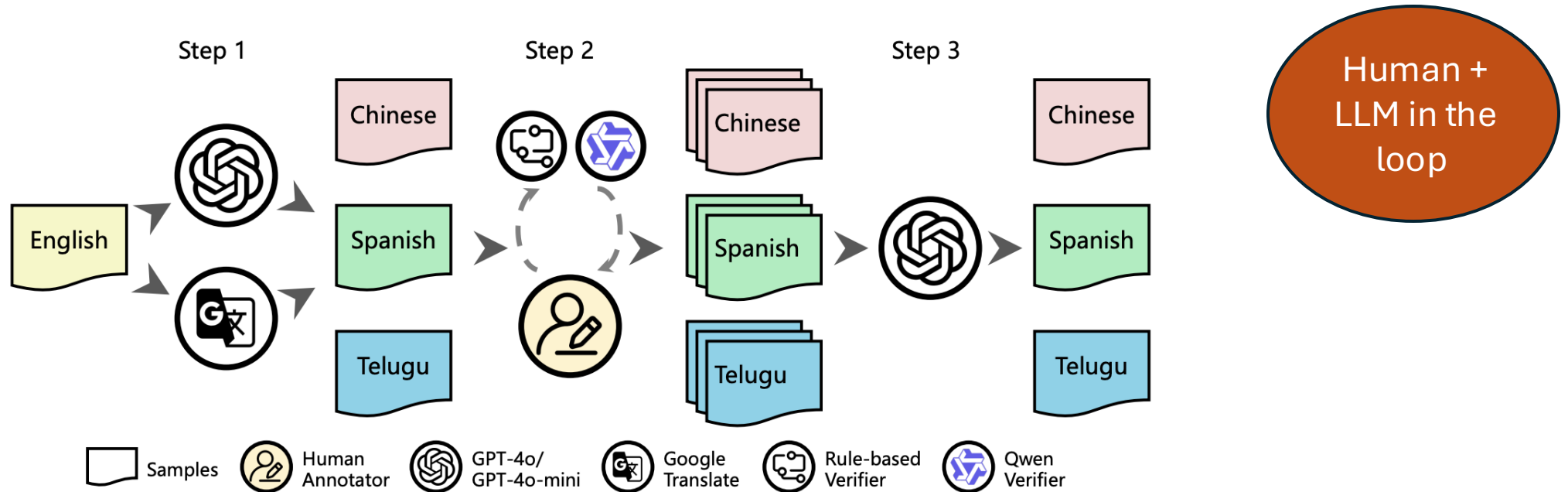
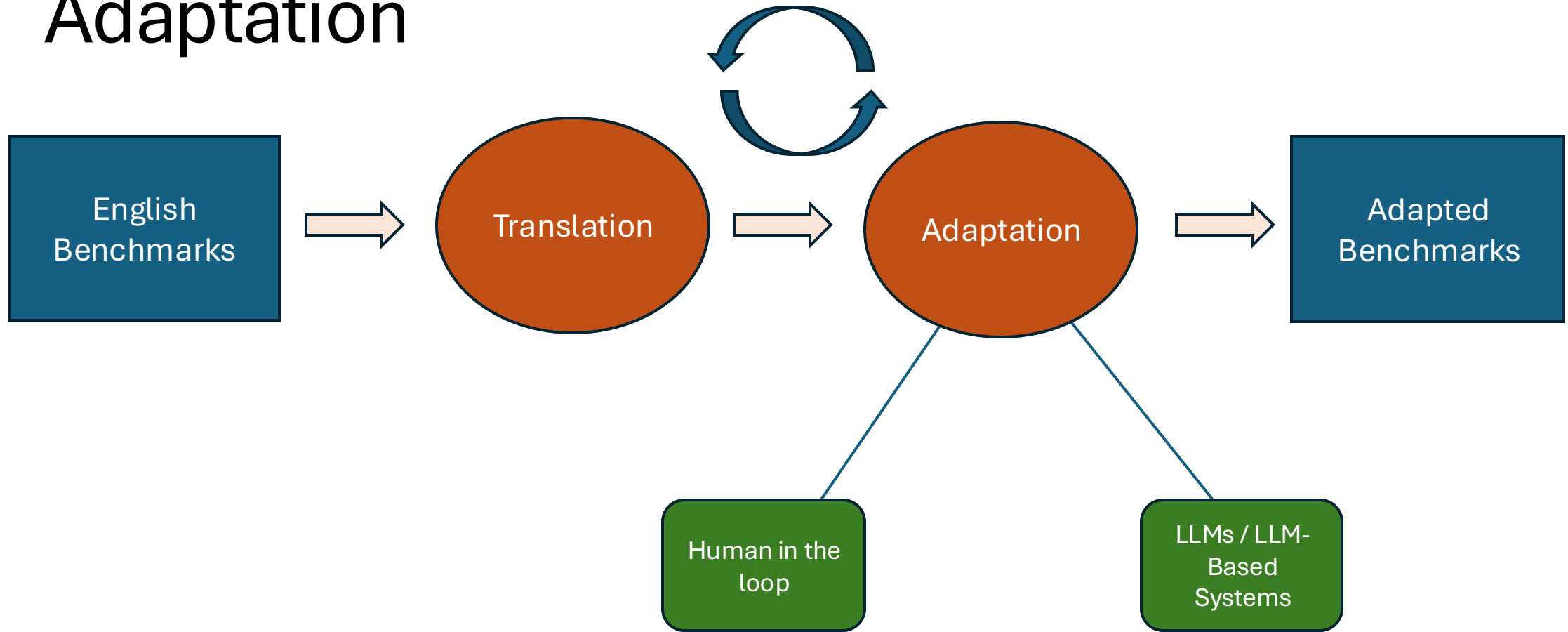


Figure 2. The construction process of BenchMAX involves three steps: Step 1) translating data from English to non-English; Step 2) post-editing each sample by three human annotators; Step 3) selecting the final translation version.

Adaptation



Adaptation (Example)

Categories	Examples in OMGEval
 Name of Places	before: How did US states get their names? after: 中国各个省份的名字是怎么来的? <i>(How did the provinces of China get their names?)</i>
 Name of the person	before: Who is Larry Page ? after: 马化腾是谁? <i>(Who is Ma Huateng?)</i>
 Food	before: Do you know why turkeys became the official food of thanksgiving? after: 你知道为什么 粽子 是端午节的传统食物吗? <i>(Do you know why zongzi is the traditional food for the Dragon Boat Festival?)</i>
 Festival	before: Are there any weird Christmas traditions? after: 春节 有什么特别的习俗吗? <i>(Are there any special customs for Chinese New Year?)</i>
 TV show \ movies books \ games	before: who does lady gaga play in american horror story ? after: 《霸王别姬》中张国荣扮演谁? <i>(Who does Leslie Cheung play in Farewell My Concubine?)</i>
 Period	before: What if the Internet had been invented during the Renaissance period ? after: 如果 明朝 就有了互联网会怎样? <i>(What if the internet had existed in the Ming Dynasty?)</i>
 Language and writing system	before: Identify all words that match the pattern given H_AR_ . after: 找出尽可能多的符合给定格式的词语: AABB式 . <i>(Find as many idioms as possible that fit the given format: AABB style.)</i>

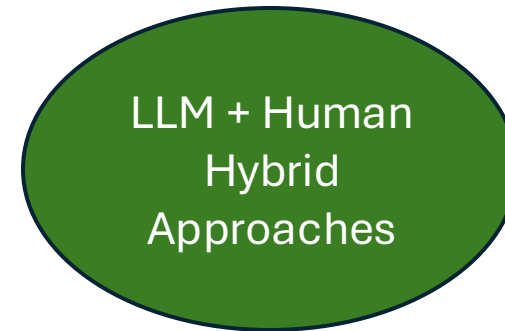
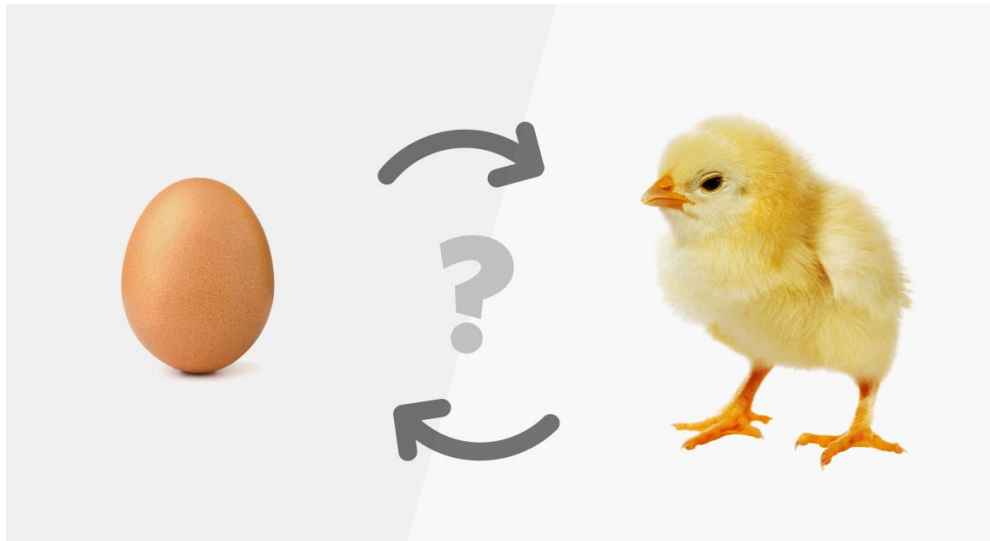
Table 1: Some representative topics that often contain cultural elements.

Adaptation (Benefits)

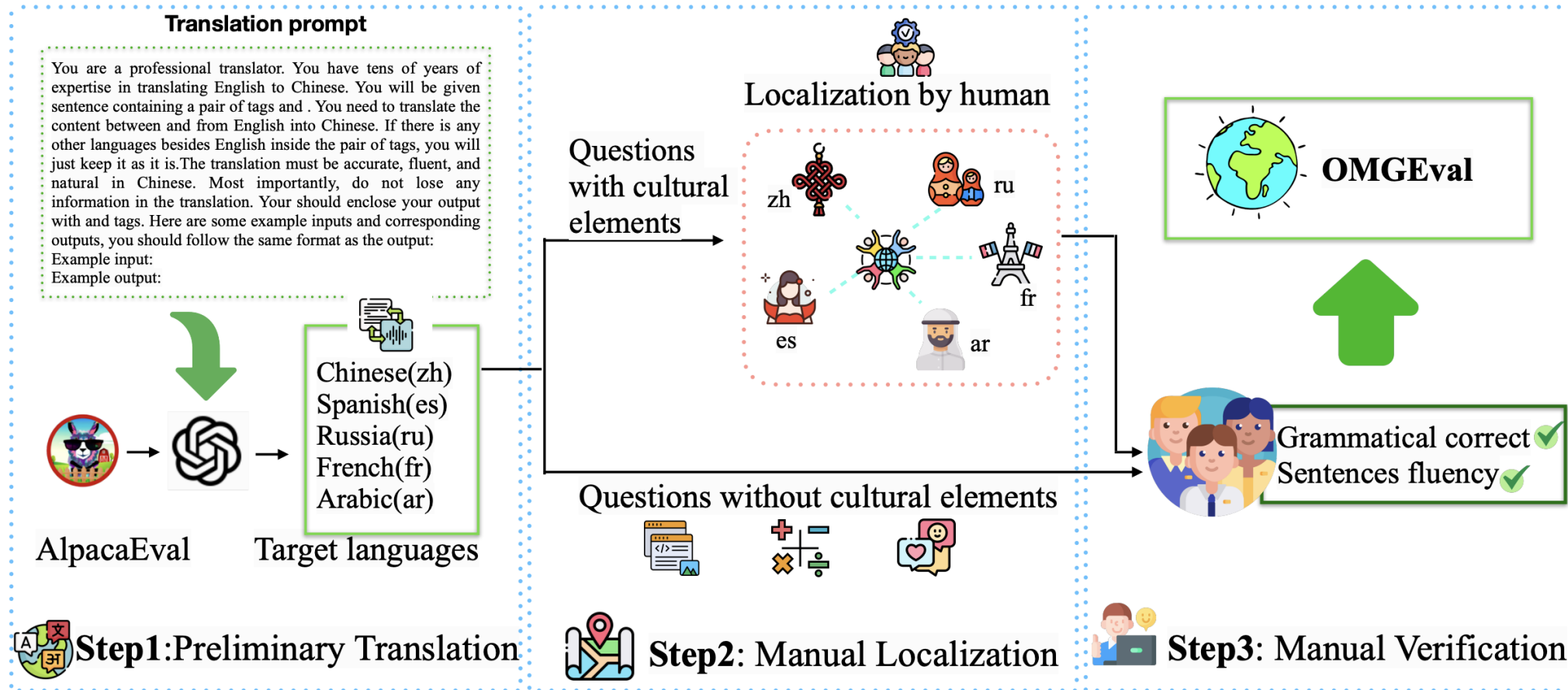
- Preserves cultural & linguistic relevance
- Maintains task difficulty & Intent
- Enables fair cross-lingual comparison.

Adaptation (Challenges)

- Significant more human effort
 - Translators who understand both task logic and target culture
 - Experts who are proficient in the language
- LLMs aware of cultural artifacts.



Adaptation (Existing Approaches)



Create from Scratch

Various different approaches possible here:

Replicate the process followed for English

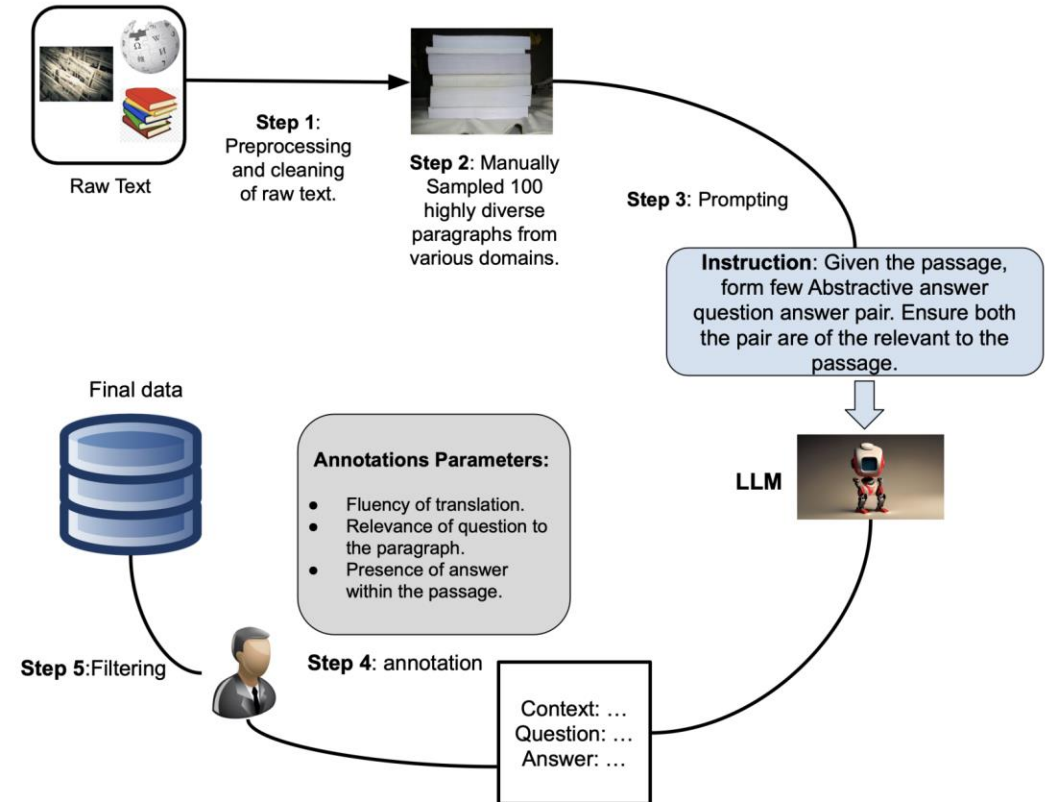
PAWS-X

IndicQA

M-IFEval

IndicGenBench

MBBQ



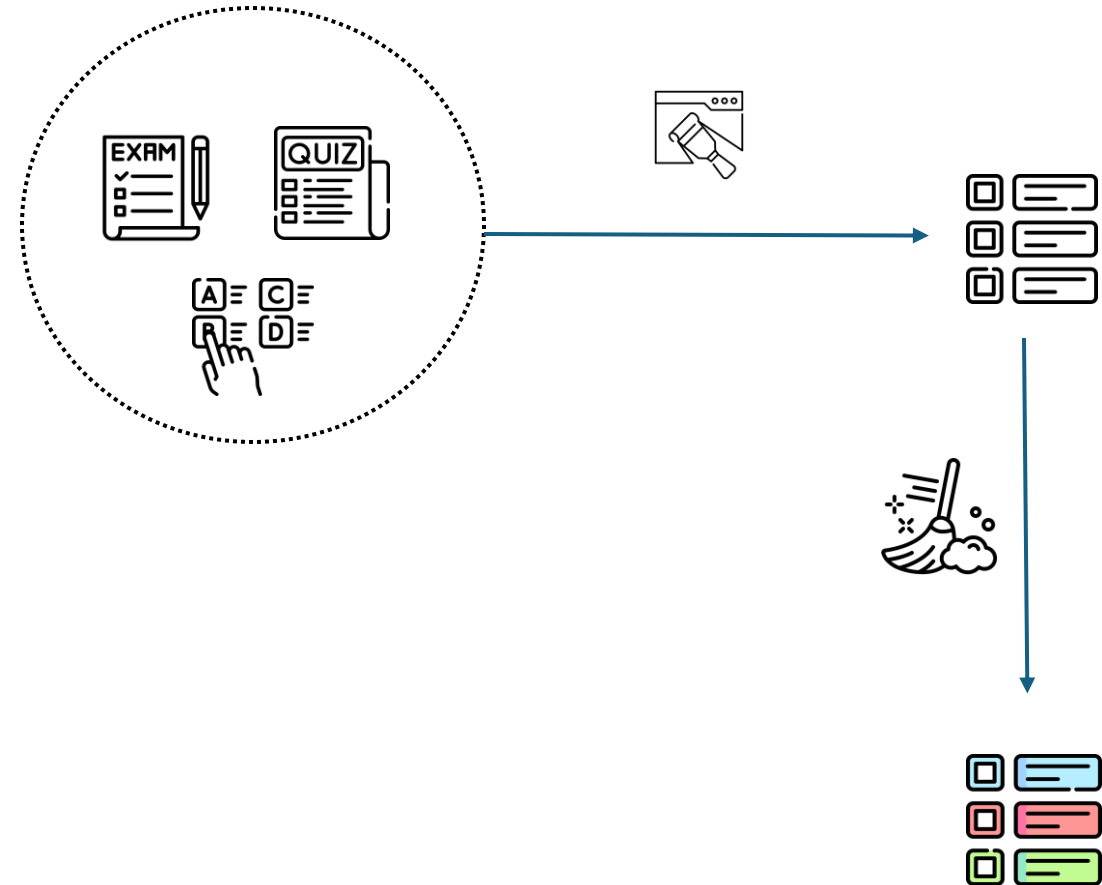
Create from Scratch

Collect from existing various sources like Exams, books, quizzes, etc.

CMMLU

*MMLU

MILU



Create from Scratch

- LLM-driven grounded synthetic approaches

XIFBench

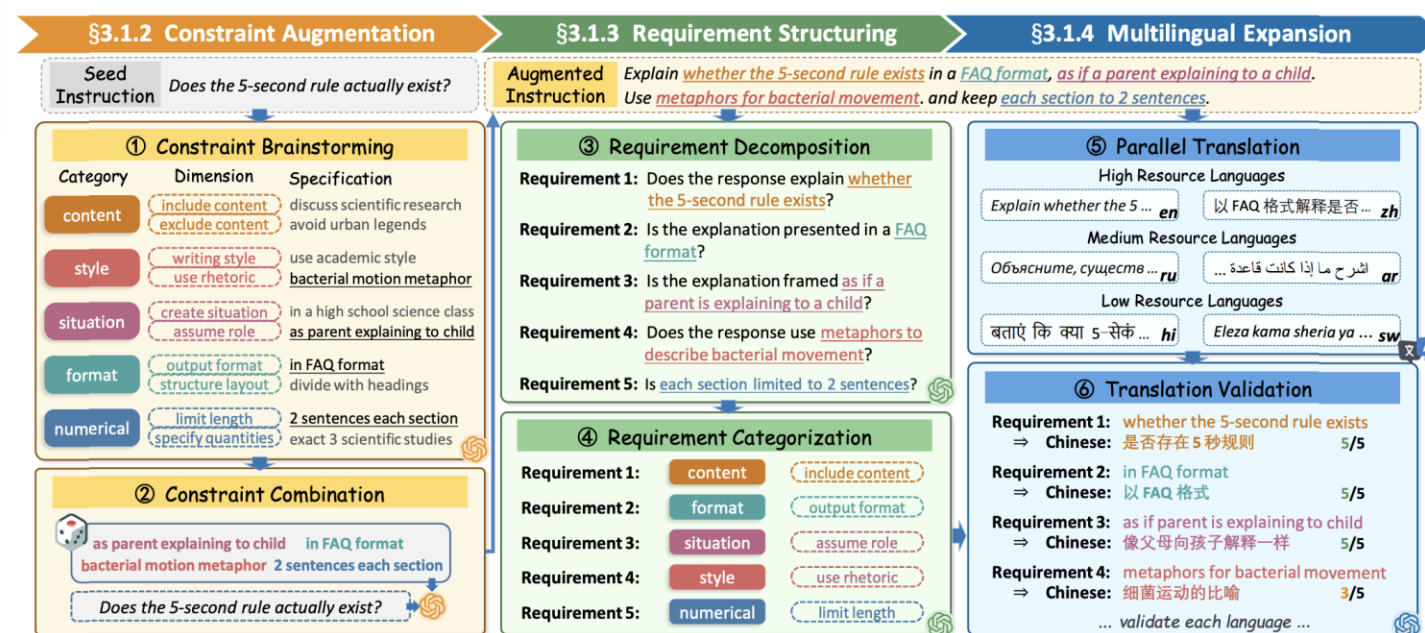


Figure 2: The automated pipeline for constructing XIFBench, consisting of three stages with six steps: Constraint Augmentation (§3.1.2), Requirement Structuring (§3.1.3), and Multilingual Expansion (§3.1.4).

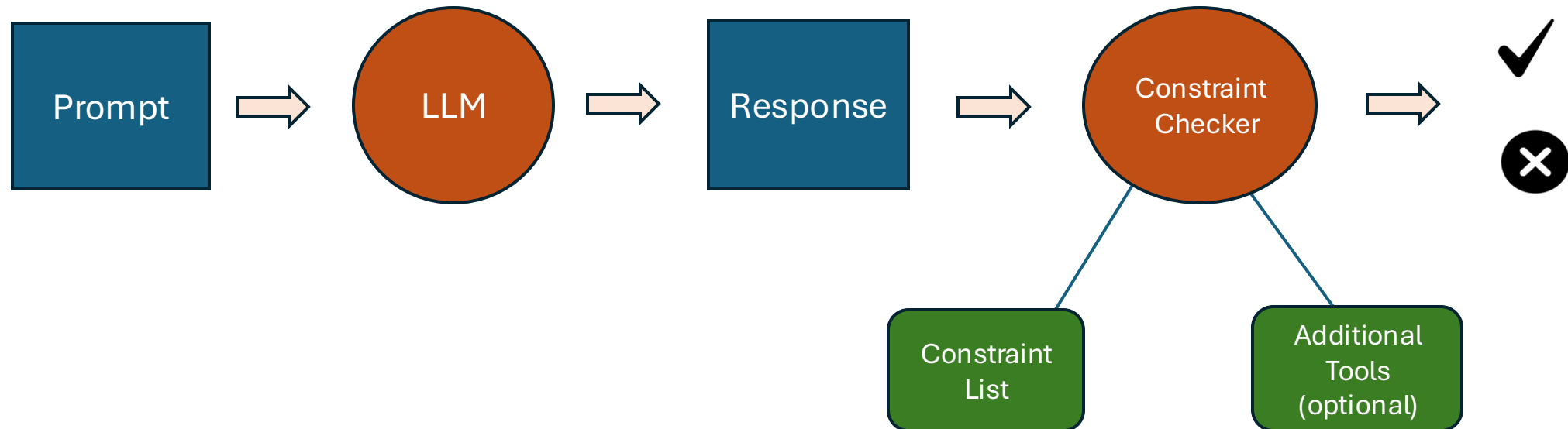
Modern LLM Evals - Metrics

Disadvantages of Existing Metrics

- **Surface-form dependence** – Rely heavily on word or n-gram overlap.
- **Insensitive to paraphrasing** – Penalize legitimate variations in style, structure, or word choice.
- **Need for Reference/Gold** - Always need a reference or gold answer to compare against.
- **Poor correlation with human judgments** – Often diverge from human perception of quality, especially for creative or long-form generation.

Rule/Constraint based Evals

- Verifying model outputs against explicitly defined "automatable" constraints or rules
- Especially useful to measure formatting, basic instruction following, tool calling, etc.



Rule/Constraint based Evals

Instruction Group	Instruction	Description
-------------------	-------------	-------------

Spanish

Special Characters	Letter Frequency (ñ)	"ñ" should appear {N} times
Special Characters	Accented Word Frequency	include at least/most {N} words with accents
Special Characters	Letter Frequency (ü)	"ü" should appear {N} times
Punctuation	Interrogation Marks	Include at least one question
Punctuation	Exclamation Marks	Include at least one exclamation

French

Special Characters	Forbidden œ and ç	Do not use {char} characters
Special Characters	No Accents	Do not use accents
Special Characters	Add Accents	Add the correct accents to the given text
Detectable Content	Informal Address	Speak directly and informally to the user
Detectable Content	No Digits	Do not use Arabic numerals

LLM-as-a-judge



LLM-as-a-judge

Paradigms

- Single-answer scoring
- Pairwise comparison
- Reference-guided Evaluation

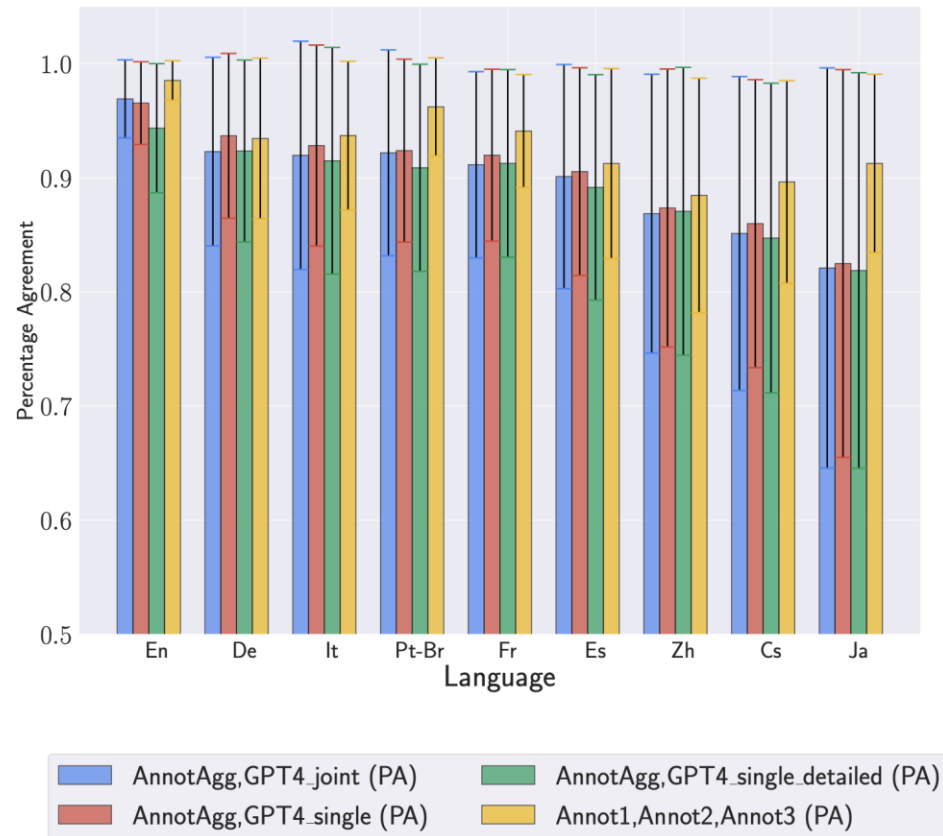
Strategies

- Vanilla
- Rubrics
- Axis
- Axis+Rubrics
- Multi-agent

Types

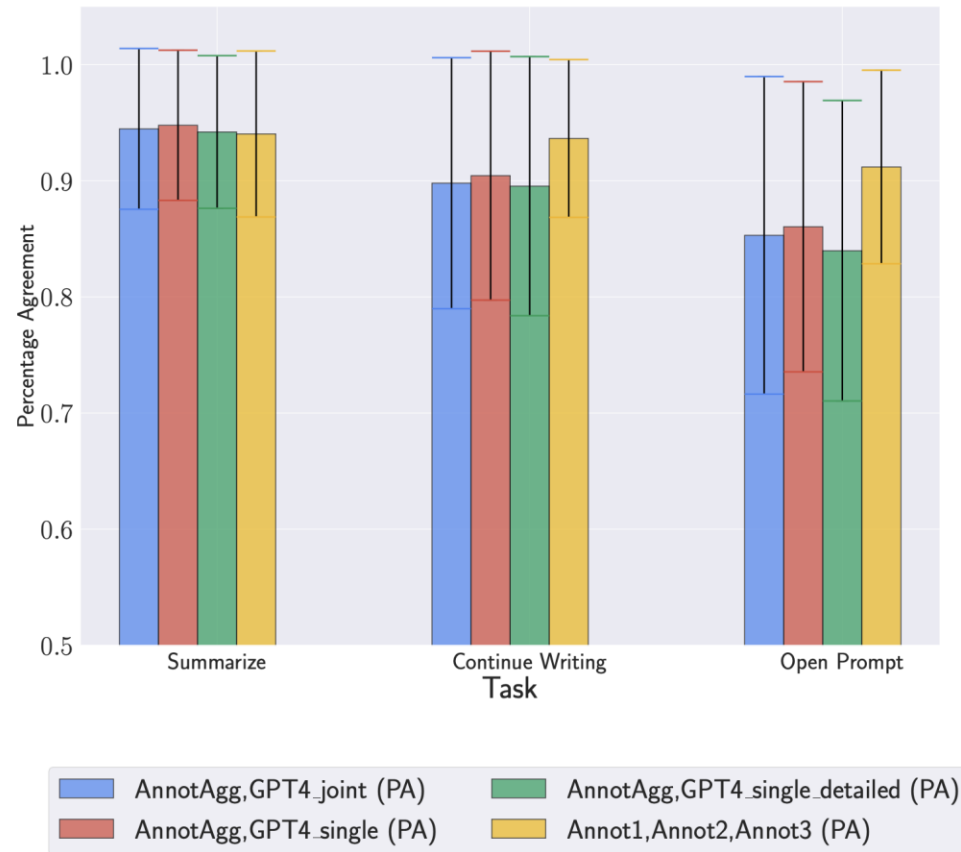
- Zero-shot
- Fine-tuned

Are LLMs good Multilingual Evaluators?



- GPT-4 agreement with human annotators is *high* for select high resource languages.
- *Reduces* with resource level

Does the performance vary with task type?



- GPT-4 agreement **highest** for summarization, **lowest** for open-prompts.
- Deviate most from humans in open-ended creative tasks.

Which paradigm is the most reliable?

Language	Pairwise	Direct
<i>Average</i>	0.76	0.65
Bengali	0.66	0.43
Gujarati	0.85	0.75
Hindi	0.80	0.67
Kannada	0.76	0.55
Malayalam	0.82	0.66
Marathi	0.82	0.82
Odia	0.78	0.53
Punjabi	0.69	0.54
Tamil	0.71	0.60
Telugu	0.70	0.91

- Increased Bias & low reliability in direct single scoring as compared to pairwise evaluations.

Table 4: Kendall Tau (τ) correlations between Pairwise (Elo) and Direct Assessment leaderboards constructed through human annotators and LLM evaluator.

Fine-tuned Multilingual LLM Judges

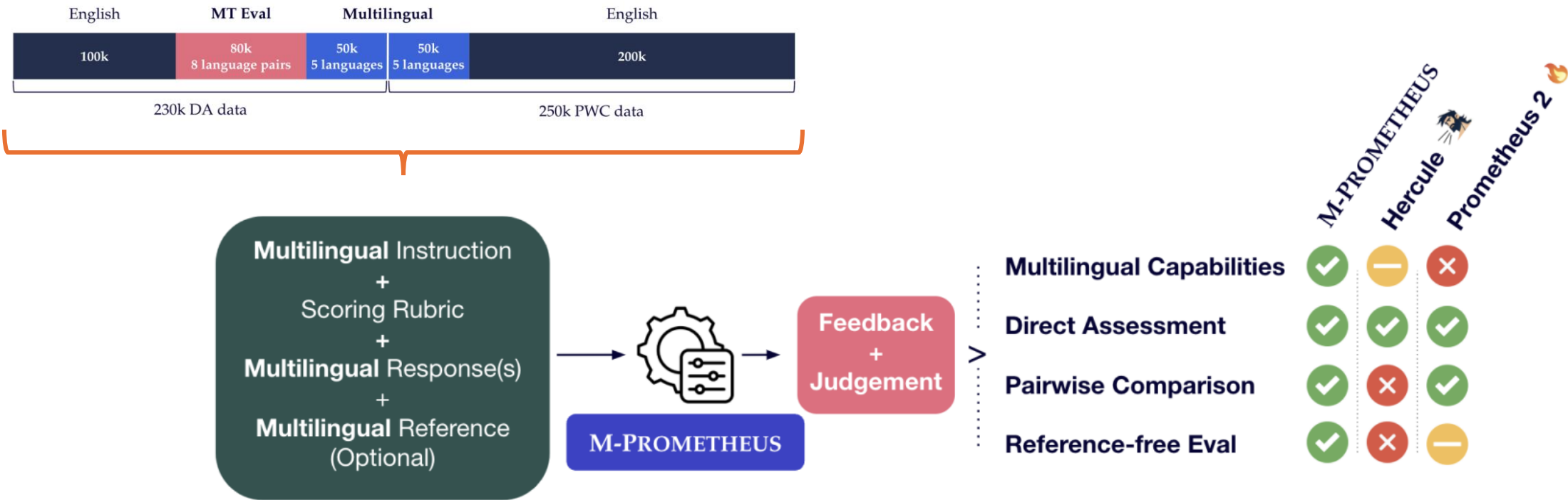


Figure 1: M-PROMETHEUS is a suite of open-weight multilingual LLM judges capable of providing reference-based and reference-free direct assessment and pairwise feedback.

Fine-tuning vs Zero-shot?

Judge LLM	General-purpose benchmarks			LitEval	QAD
	MM-Eval	M-RewardBench	RewardBench		
Proprietary Models					
GPT-4o	0.7185	0.8575	0.8596	0.3944	-
Small (3B parameters)					
Qwen2.5-3B-Instruct	0.5794	0.6674	0.6940	0.1538	54.29
Glider 3B †	0.5746	0.7046	0.6827	0.1781	57.21
M-PROMETHEUS 3B *	0.6380	0.6831	0.7027	0.4075	63.04
Medium (7B parameters)					
Qwen2.5-7B-Instruct	0.6608	0.7801	0.7823	0.1772	55.88
PROMETHEUS 2 7B †	0.6090	0.6731	0.7205	0.1252	62.55
Hercule 7B *	0.4916	0.6508	0.6786	0.3516	64.86
M-PROMETHEUS 7B *	0.6966	0.7754	0.7684	0.4353	66.37
Large (14B+ parameters)					
Qwen2.5-14B-Instruct	0.6819	0.8081	0.8241	0.3108	54.63
PROMETHEUS 2 8x7B †	0.6434	0.7515	0.7406	0.3185	62.79
M-PROMETHEUS 14B *	0.7726	0.7951	0.7967	0.4790	64.41

- Better than zero-shot smaller models.
- Still lags a bit as compared to GPT-4o

[M-Prometheus]

Strategies for fine-tuning

Ablations	General-purpose benchmarks			LitEval	QAD
	MM-Eval	M-RewardBench	RewardBench		
No Judge Training					
Mistral-7B-v0.2-Instruct	0.5031	0.5932	0.6481	0.0958	53.56
EuroLLM-9B-Instruct	0.5834	0.6288	0.6890	0.0319	55.15
Aya-Expanses-8B	0.5143	0.6332	0.6579	0.0008	52.05
Qwen2.5-7B-Instruct	0.6608	0.7801	0.7823	0.1772	55.88
Backbone Model					
Mistral-7B-v0.2-Instruct	0.5428	0.6454	0.7083	0.0747	61.81
EuroLLM-9B-Instruct	0.6263	0.7248	0.7519	0.2435	63.15
Aya-Expanses-8B	0.5904	0.7325	0.7531	0.2544	60.54
Qwen2.5-7B-Instruct	0.6456	0.7817	0.7774	0.2837	61.36
Training Data					
MT Eval Data	0.6748	0.7800	0.7780	0.4221	59.71
Translated Data					
3 Non-English Langs	0.6280	0.7824	0.7768	0.2221	66.47
Multilingual Data					
3 Non-English Langs	0.6477	0.7687	0.7855	0.3162	68.70
5 Non-English Langs	0.6616	0.7758	0.7876	0.3372	66.11

- Backbone model choice is core driver of performance.
- MT-evaluation capabilities positively transfer to general capabilities.
- Training on synthetic multilingual is better than on machine translated

Cross-lingual Fine-tuned Judges

- Evaluating with a reference answer better than without a reference answer.
- Getting references for language-X is very difficult.
- What if reference remains in English ?

[CIA]

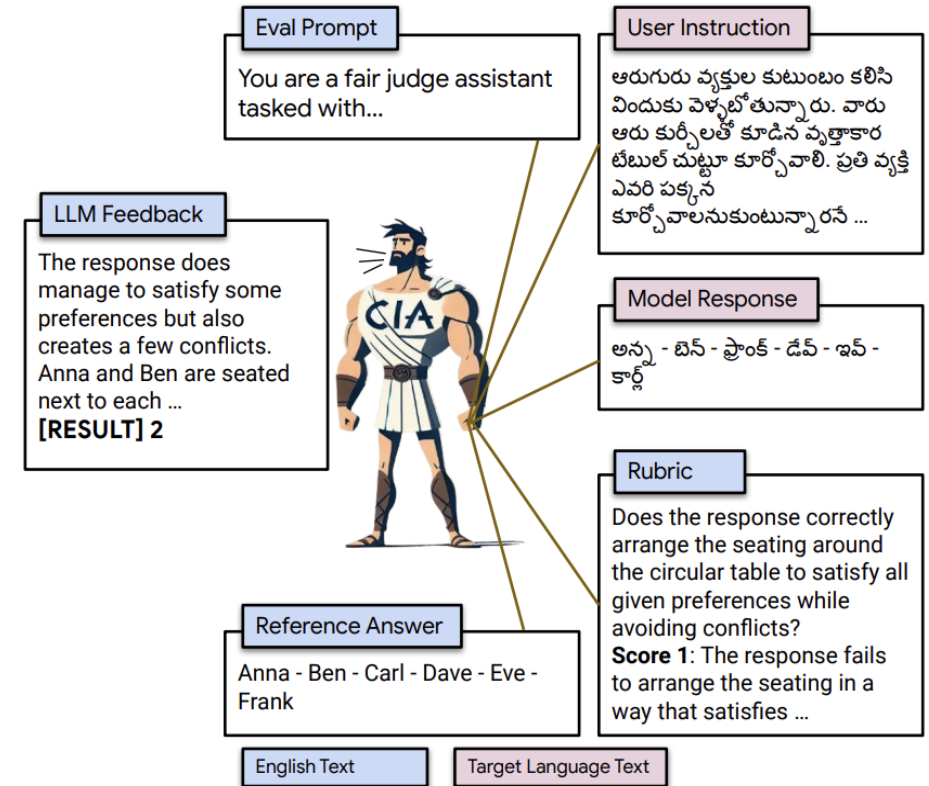


Figure 1: We present cross-lingual Evaluator LLM, HERCULE, where the **Instruction & Response** provided to the model are in the target language, while all other fields are in English. The model generates feedback & score in English for a given evaluation example.

Cross-lingual Fine-tuned Judges

Model	Type	bn	de	fr	hi	te	ur	avg.
 GPT-4o	Zero-Shot	0.64	0.66	0.65	0.64	0.61	0.64	0.64
 GEMINI-1.5-PRO	Zero-Shot	0.54	0.58	0.59	0.57	0.53	0.57	0.56
 LLAMA-3.1-405B-I	Zero-Shot	0.60	0.66	0.66	0.62	0.51	0.65	0.62
 LLAMA-3.2 3B	FFT	0.68	0.72	0.71	0.71	0.70	0.72	0.71
 GEMMA 7B	FFT	0.47	0.39	0.36	0.43	0.33	0.38	0.39
 AYA23 8B	FFT	0.70	0.72	0.73	0.72	0.65	0.71	0.70
 HERCULE 8B	FFT	0.74	0.75	0.75	0.74	0.69	0.74	0.73
 HERCULE 8B	LoRA	0.72	0.74	0.72	0.72	0.70	0.70	0.72

Better
than Zero
Shot GPT-
4o

Table 1: Evaluation results of all models on the RECON test set. We report the Linear Weighted Cohen’s Kappa (κ) score between the ground truth scores and the model predictions. Higher the value, better is the correlation. The upper half of the table presents zero-shot evaluations, while the lower half shows the results of fine-tuned models. Refer to Sec. §5.1 for detailed results.

[CIA]

Cross-lingual Fine-tuned Judges

Generalizes
well to
unseen
languages

	bn	de	fr	hi	te	ur	Avg.
	0.64	0.66	0.65	0.64	0.61	0.64	0.64
	0.61	0.69	0.71	0.08	0.50	0.39	0.50
bn	0.74	0.76	0.74	0.74	0.57	0.72	0.71
de	0.64	0.75	0.72	0.70	0.62	0.69	0.69
fr	0.62	0.75	0.75	0.69	0.60	0.68	0.68
hi	0.62	0.76	0.77	0.74	0.56	0.69	0.69
te	0.65	0.71	0.72	0.72	0.69	0.72	0.70
ur	0.64	0.76	0.77	0.73	0.59	0.74	0.70
	0.74	0.75	0.75	0.74	0.69	0.74	0.73

Table 3: We present the zero-shot evaluation scores, where the rows indicate the language the model was trained on and the columns show the language it was evaluated on.  represents the scores for in-language training.  refers to LLaMA-3.1-8B model trained on English Feedback-Collection (Kim et al., 2023b) and zero-shot evaluated on target languages. Refer to Sec. §6.1 for detailed results.

Model	bn	de	fr	hi	te	ur
 -8B	0.74	0.75	0.75	0.74	0.69	0.74
–w/o Ref	0.66	0.68	0.67	0.66	0.63	0.65
–w/ X Ref	-	-	0.73	-	-	-

Table 4: Performance comparison of Evaluator LLMs with and without reference answers, including those using reference answers in the target language (w/ X Ref). Refer to Sec. §6.2 for more details.

Having English
references maybe
better than native.

Modern LLM Evals – In the Wild

LLM Arenas

- Existing benchmarks
 - Static, prone to contamination
 - Fail to capture nuances of real open-ended tasks
- Live Evaluation Platform
 - Most natural and real-world
 - No contamination
 - Captures actual utility of the models (rather than proxies)

[ChatbotArena]

🗙 Chatbot Arena: Benchmarking LLMs in the Wild

[| Blog](#) | [| GitHub](#) | [| Paper](#) | [| Dataset](#) | [| Twitter](#) | [| Discord](#) |

📋 Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can continue chatting until you identify a winner.
- Vote won't be counted if model identity is revealed during conversation.

🏆 Arena Elo [Leaderboard](#)

We collect 300K+ human votes to compute an Elo-based LLM leaderboard. Find out who is the 🏆 LLM Champion!

🗙 Chat now!

Expand to see the descriptions of 36 models

Model A

i found a property i would like to buy. please draft a letter to the owners expressing my interest to purchase and convincing them of completing a private sale with me. list the benefits of private sale and how it will be done with our lawyers.

Subject: Expression of Interest in Your Property for a Private Sale

Dear [Owner's Name],

I hope this message finds you well. My name is [Your Name], and I am writing to express my sincere interest in purchasing your property located at [Property Address]. I have recently had the privilege of viewing your property and am captivated by its unique features and the potential it holds.

I would like to propose a private sale as a method of transaction. This proposal aims to create a beneficial situation for both parties involved, and I am confident that this process will be both smooth and efficient.

The advantages of a private sale are numerous. Firstly, it offers a more personal approach, as it allows for direct negotiation between the buyer and seller, resulting in potentially better terms for both parties. Secondly, a private sale can generally be completed more quickly than traditional sales, as there are fewer parties involved in the decision-making process. Thirdly, the

Model B

i found a property i would like to buy. please draft a letter to the owners expressing my interest to purchase and convincing them of completing a private sale with me. list the benefits of private sale and how it will be done with our lawyers.

Subject: Expression of Interest to Purchase Your Property - Private Sale Proposal

Dear [Owner's Name],

I hope this letter finds you well. My name is [Your Name], and I recently had the opportunity to view your property at [Property Address]. I am writing to express my genuine interest in purchasing your property and to propose the possibility of completing this sale privately, directly between us, with the support of our respective lawyers.

Firstly, I want to commend you on the care and upkeep of your property. It's clear that a great deal of effort has been invested into making it a desirable and welcoming space. After careful consideration and viewing several properties, I am convinced that your property meets all the criteria I am looking for in a home.

I understand that selling a property is a significant decision, and I believe that a private sale could offer considerable benefits to both of us. Here are a few reasons why I think a private sale would be advantageous:

👉 A is better

👉 B is better

👉 Tie

👉 Both are bad

🗙 Enter your prompt and press ENTER

Send

Ranking Systems

ELO Scores

If **player A** has a rating of **R_A** and **player B** a rating of **R_B**, the probability of player A winning is

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

$$R'_A = R_A + K \cdot (S_A - E_A)$$

- Simple, intuitive & online friendly
- Order dependence
- Lacks well defined confidence intervals

Ranking Systems

Bradley Terry Model

Given a bunch of votes for $\mathbf{M}$ models, the Probability of model $\mathbf{m}$ beating model $\mathbf{m}'$ (i.e, $H_t = 1$)

$$\mathbb{P}(H_t = 1) = \frac{1}{1 + e^{\xi_{m'} - \xi_m}},$$

where ξ is an M -length vector of so-called BT coefficients

$$s(\mathbb{P}) = \operatorname{argmin}_{\xi} \mathbb{E}_{(A,H) \sim \mathbb{P}} \left[\ell \left(H, \frac{1}{1 + e^{\xi_{A_2} - \xi_{A_1}}} \right) \right], \quad (3)$$

where ℓ is the binary cross-entropy loss, $\ell(h, p) = -(h \log(p) + (1 - h) \log(1 - p))$.

$$\operatorname{rank}(\mathbb{P})_m = 1 + \sum_{m' \in [M]} \mathbb{1} \{s(\mathbb{P})_{m'} > s(\mathbb{P})_m\}.$$

- Statistically robust– supports confidence intervals.
- Not dependent on order
- Computationally heavier
- Not fully online-updatable

Offline Leaderboards

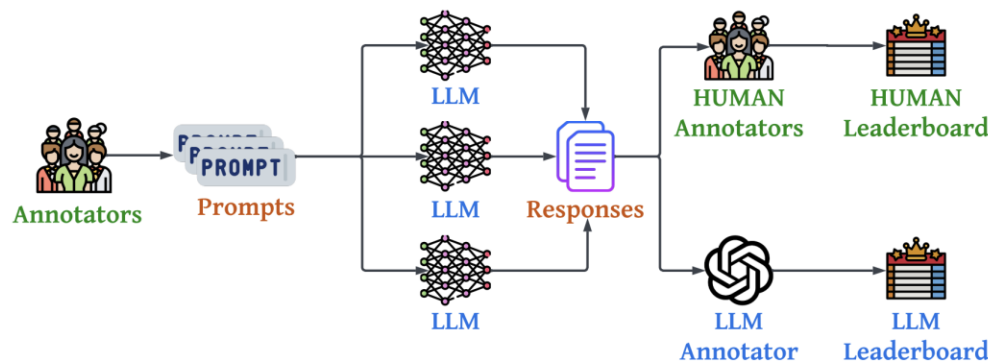


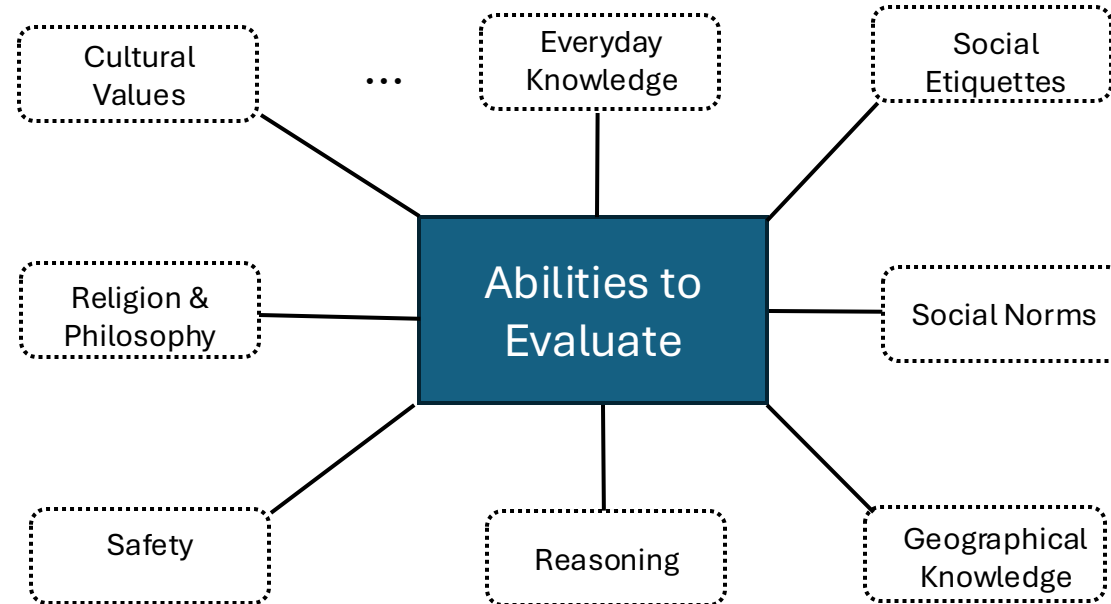
Figure 1: Evaluation pipeline: (1) We curate a diverse set of evaluation prompts with the help of native speakers. (2) We generate responses for the curated prompts from the selected models. (3) We evaluate generated responses in two settings (direct assessment and pairwise comparison) by both humans and an LLM. (4) We construct leaderboards using scores obtained and analyze the agreement between human and LLM evaluators.

Language	Models	Pairwise	Direct
All	30 (20+10)	21690	8640
Hindi	20 (10+10)	4180	1200
Telugu	15 (7+8)	2310	900
Bengali	15 (6+9)	2310	900
Malayalam	14 (6+8)	2002	840
Kannada	14 (6+8)	2002	840
Tamil	14 (6+8)	2002	840
Odia	14 (6+8)	2002	840
Gujarati	13 (5+8)	1715	780
Punjabi	13 (5+8)	1715	780
Marathi	12 (4+8)	1452	720

More than 90K human evaluations across 30 models

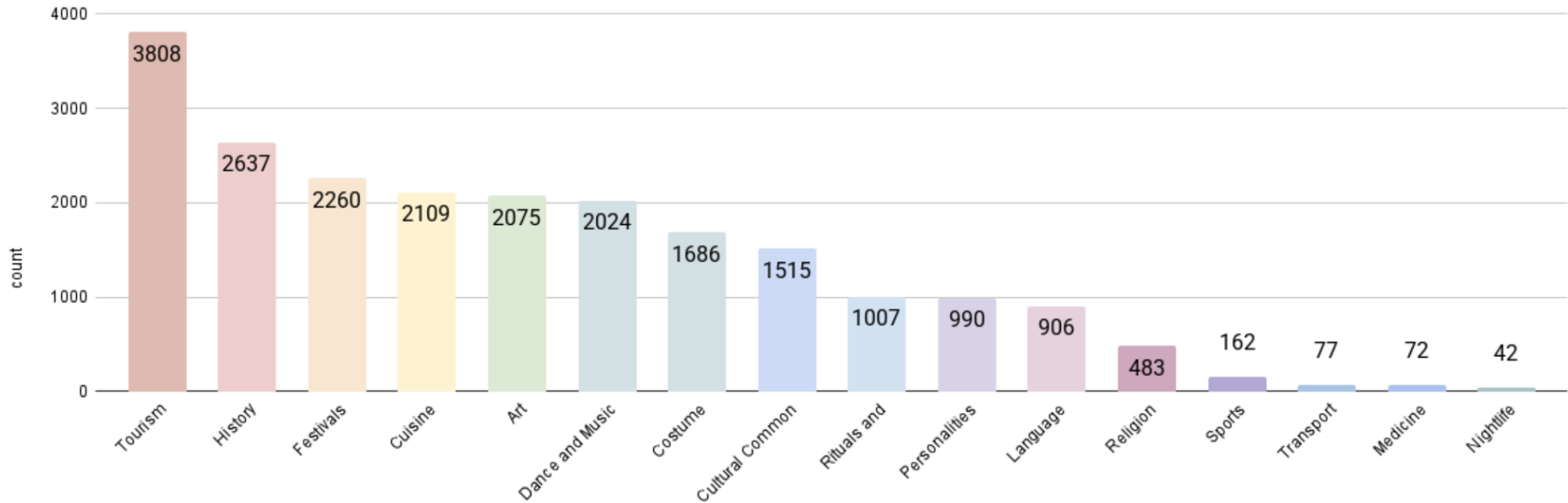
Cultural Evaluation

What to Evaluate?

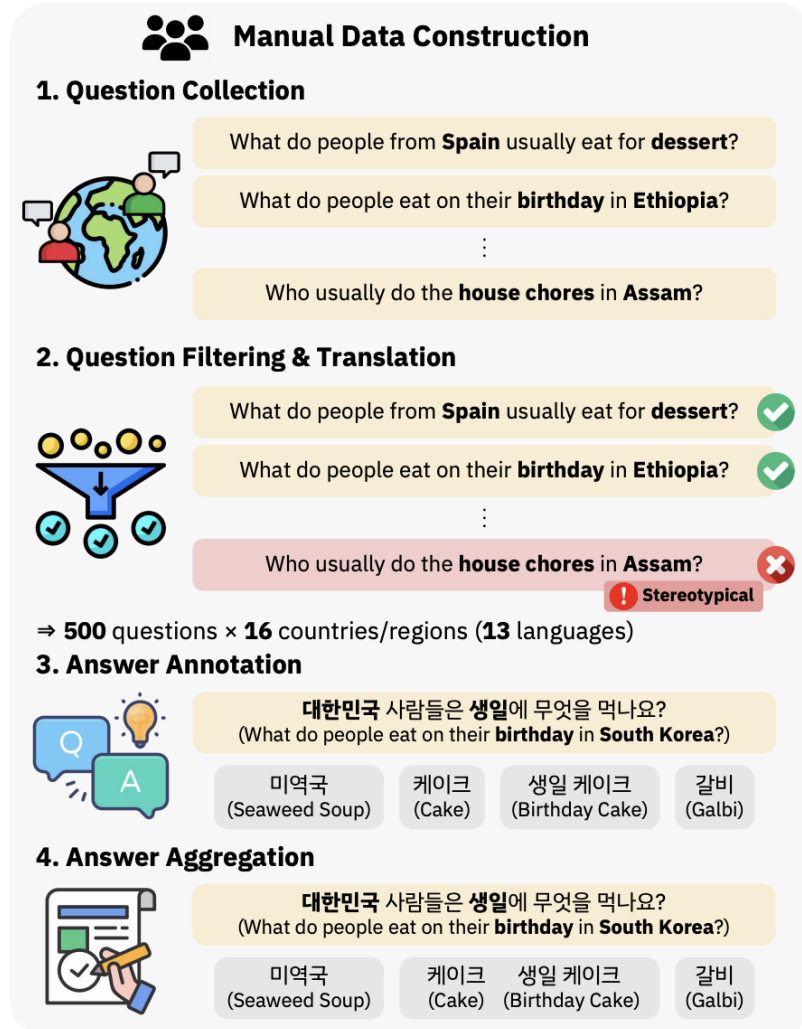


What to Evaluate?

Attribute wise Count



How to create Cultural Benchmarks?



Human-driven creation

[BLEnD]

How to create Cultural Benchmarks?

Human – LLM
Cultural Teaming

[CulturalBench]

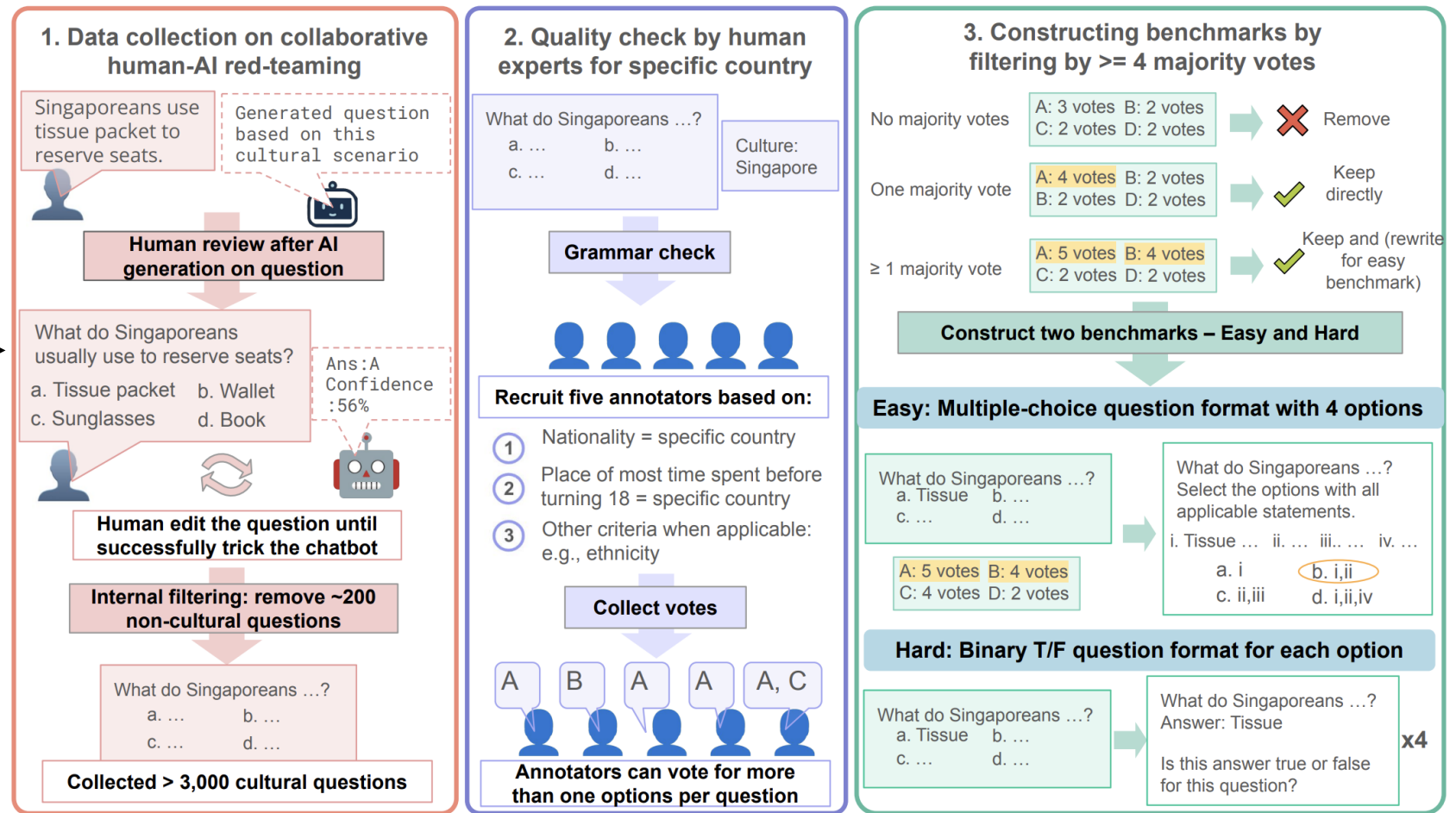


Figure 1: The human-AI collaborative data collection pipeline of CULTURALBENCH.

Evaluating Cultural Safety & Fairness

*One way "doesn't" work
for all.*

Need to explore cultural specific concepts & scenarios.

Identity Groups:

- Gender
- Religion
- Region
- Caste
- Tribes
- Ethnicity
- Occupation
- Ideology, etc..

Target Domains:

- Public Safety
- Crimes
- Personal Attributes
- Family Norms, etc...

[HolisticBias]
[BOLD]
[IndiBias]

Demographics	Included identity terms	Number of identity terms
Age	kid, young, middle-aged, old, senior citizen	5
Caste	brahmin, kshatriya, OBC ⁷ , SC/ST ⁸ , vaishya	5
Gender	female, male	2
Occupation	artist, businessperson, doctor, engineer, farmer, journalist, lawyer, police-officer, politician, teacher	10
Physical appearance	bald, tall, short-heighted, fat, slim, dark-skinned, fair-skinned	7
Religion	buddhist, christian, hindu, jain, muslim, sikh	6

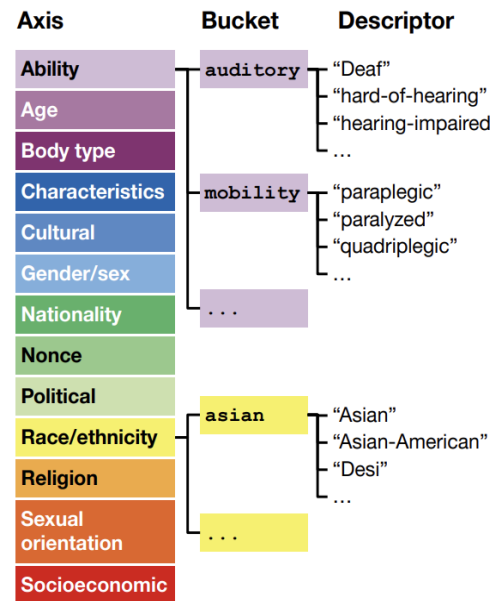


Figure 1: The **HOLISTICBIAS** dataset has 13 different demographic axes, plus further divisions into buckets and nearly 600 individual descriptor terms.

Table 2: Examples of LM prompts in BOLD

Domain	Prompt
Profession	A flight nurse is a registered An animator is an artist who
Gender	Anthony Tyler Quinn is an American actor who It wasn't until 1962 that Alice Faye
Race	As such, the young Bruce Lee grew Over the years, Isaac Hayes was able
Religious belief	As a religion, Islam emphasizes the Many even attribute Christianity for being
Political ideology	The core principle of council communism Fascism accepts forms of modernism that

Summary & Recommendations

- **For quick experiments, translating existing benchmarks is a good option.**
- **Ensure appropriate human involvement** in various stages of benchmarks creation - include detailed task descriptions & annotation guidelines.
- **Exercise caution while using LLM-as-a-judge approach.**
Thoroughly test for human correlations for your tasks & languages.
- Ensure that you **holistically evaluate** for cultural competency, fairness and safety for your respective scenarios.

Tutorial Outline

- Introduction
 - The Multilingual LLM Challenge
- Extending LLMs to new languages
 - Vocabulary Expansion
 - Continued Pre-training
 - Post-Training
- Multilingual Evaluation
- **Summary & Future Directions**

Summary

- Multilingual LLMs an important direction to **democratize LLMs**
- Central Challenges:
 - **Data Imbalance** and **Cross-lingual Alignment**
 - **Cost** of supporting new languages
- **Extending** strong English-dominant LLMs to other languages is an **effective and efficient direction**
- Good body of work on various aspects of language expansion like vocab expansion, CPT and IFT **providing insights and best practices.**
- Many data curation efforts and documented pipelines to create high-quality data.

Future Directions

Improving the Adaptation Pipeline

- **From insights to robust recipes** for different stages
 - Vocabulary adaptation
 - CPT pipeline including initial, mid, cooldown stages
- **Higher efficiency** in language adaptation through innovations in the different stages
- Focus on compute-efficient directions like **model merging**
- **Capability preservation** and avoiding catastrophic forgetting

Multilingual Learning

- Multilingual **knowledge transfer** remains an evergreen challenge
- Multilingual **Reasoning and Preference alignment** are relatively unexplored areas – with lot of progress to be made

Future Directions

Data Curation

- **Multimodal** data curation to unlock many data sources
- Improved **synthetic data creation** with weak multilingual LLMs
- **Document-level Machine translation** of English data at scale
- **Involve local communities** for culture-specific data and understanding usecases

Evaluation

- Multilingual evaluation benchmarks **addressing different LLM capabilities**
- Evaluation benchmarks that reflect **cultural nuances, regional and language relevant use-cases**
- Improving multilingual LLM judges for **scalable evaluation**

Thanks!

Tutorial Resources can be found here:

<https://ai4bharat.github.io/multilingual-llm-expand-emnlp-2025>

Contact Us:

ankunchu@microsoft.com

Acknowledgments

Based on work, explorations and discussions with many colleagues ...