Research and Challenges of Multilingual Large Language Models

Shujian Huang & Wenhao Zhu
huangsj@nju.edu.cn, zhuwh@smail.nju.edu.cn
National Key Laboratory of Novel Software Technology
School of Computer Science, Nanjing University

Tutorial Roadmap

- Chapter I: Background
- Chapter II: Observations and Analyses
- Chapter III: Enhancing LLM for More Languages
- Chapter IV: Aligning Non-English to English
- Chapter V: Future Challenges

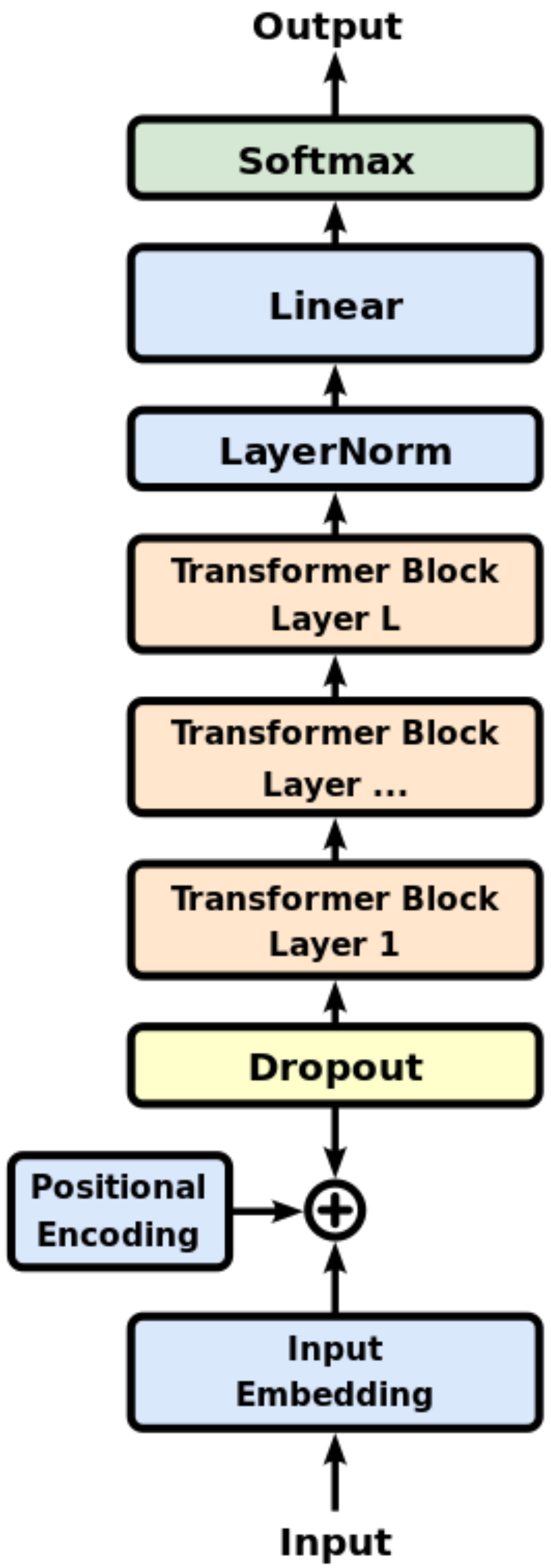
Language Model

- Language modeling aims at predicting the probability of the next token w_t based on the prefix $p(w_t|w_1 w_2 \cdots w_{t-1})$:

Hangzhou is a __

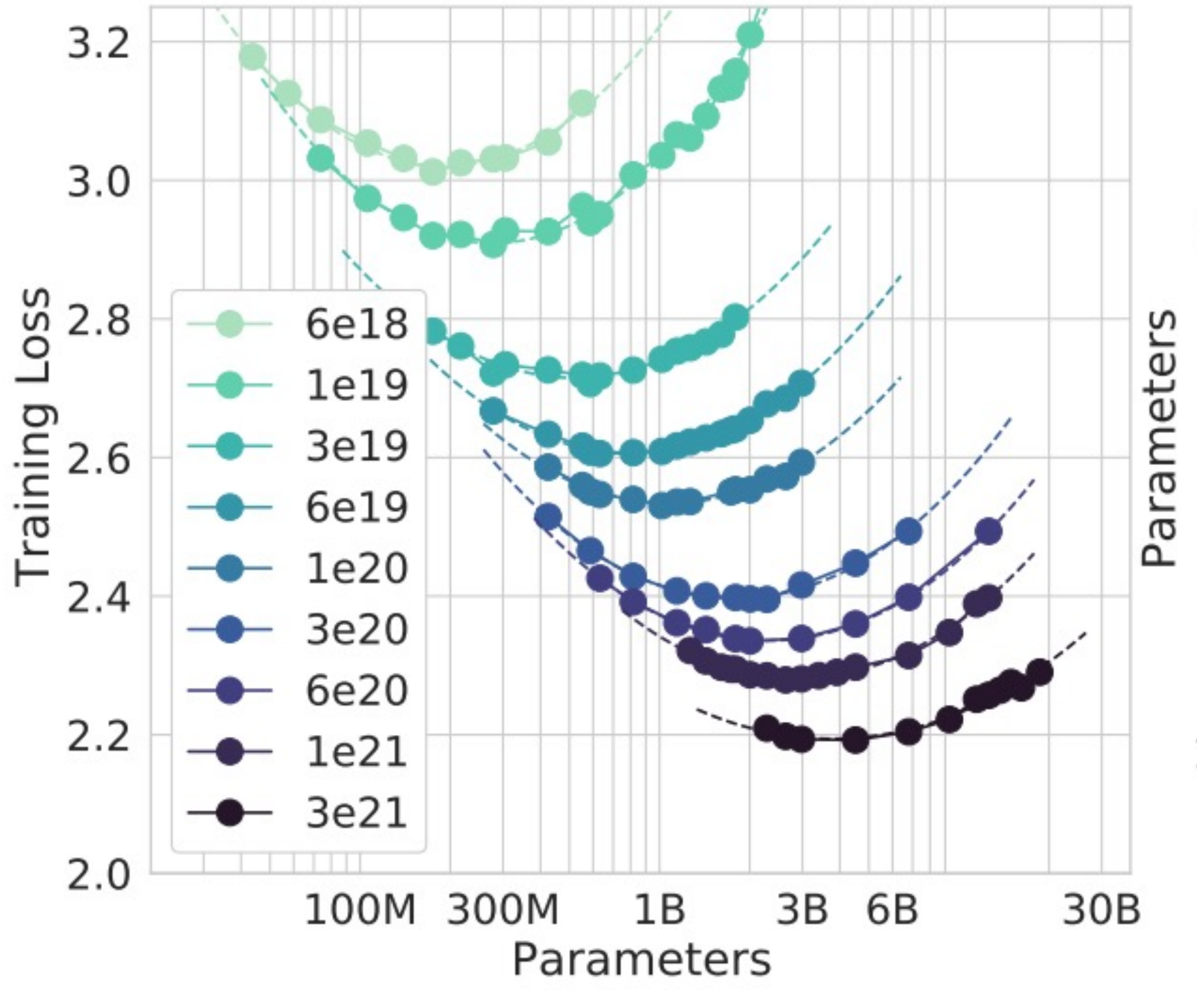
city	70%
beautiful	10%
…	
island	0.001%

- Architecture
  - Transformer has become the backbone architecture of language model.

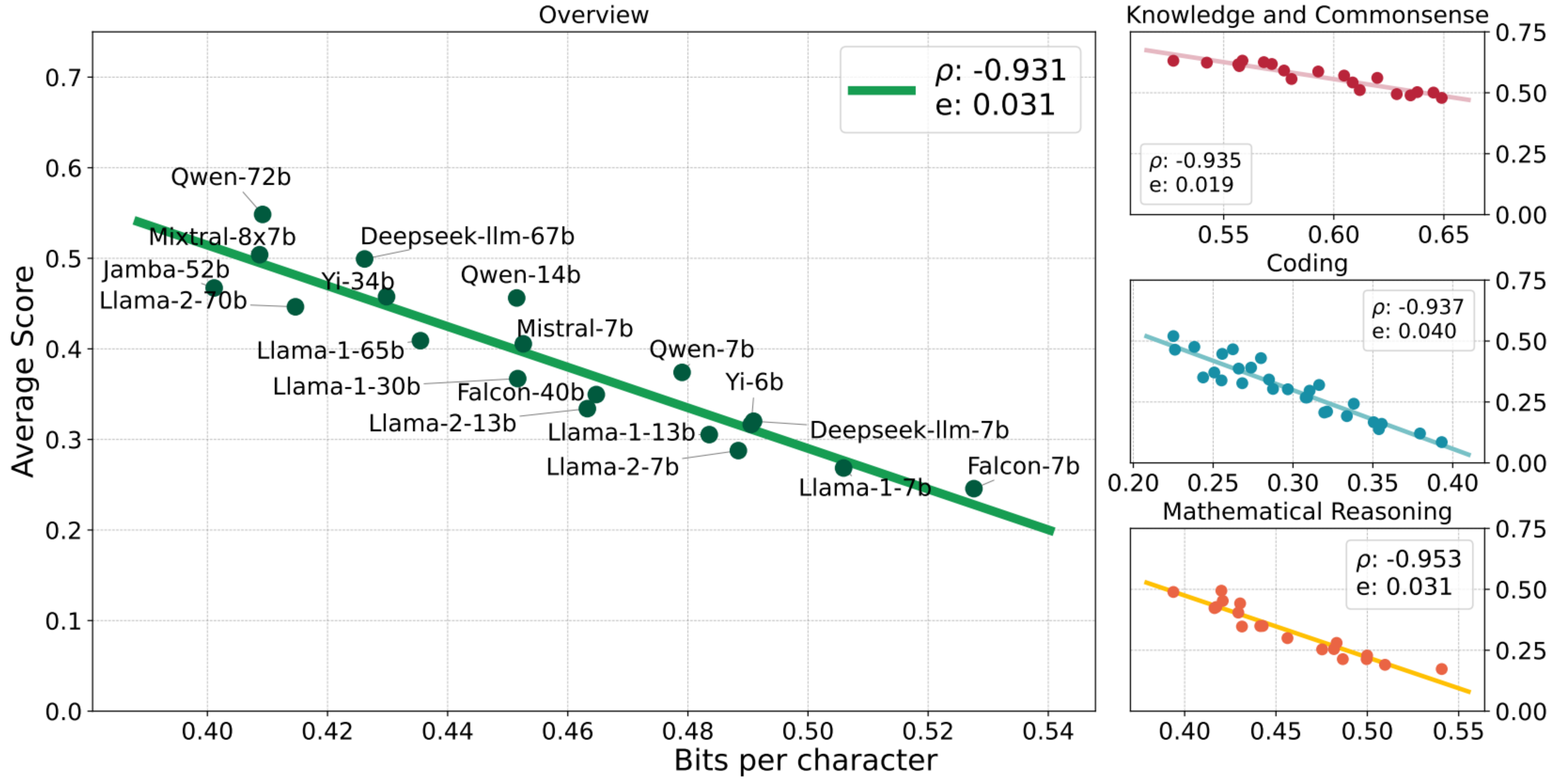


Scaling-up Language Model

- Increasing training tokens and parameters leads to lower training loss (higher-level intelligence).



$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

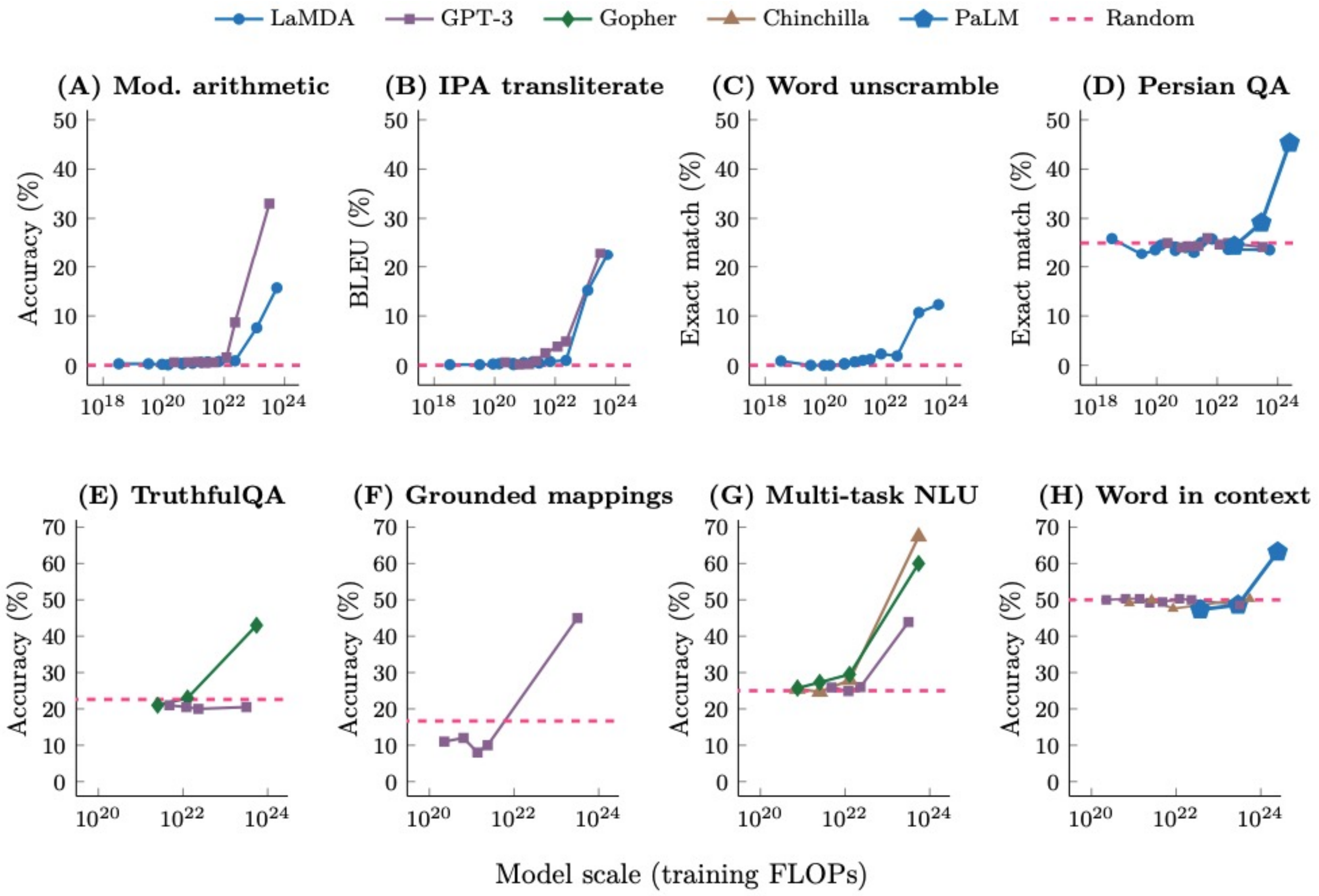


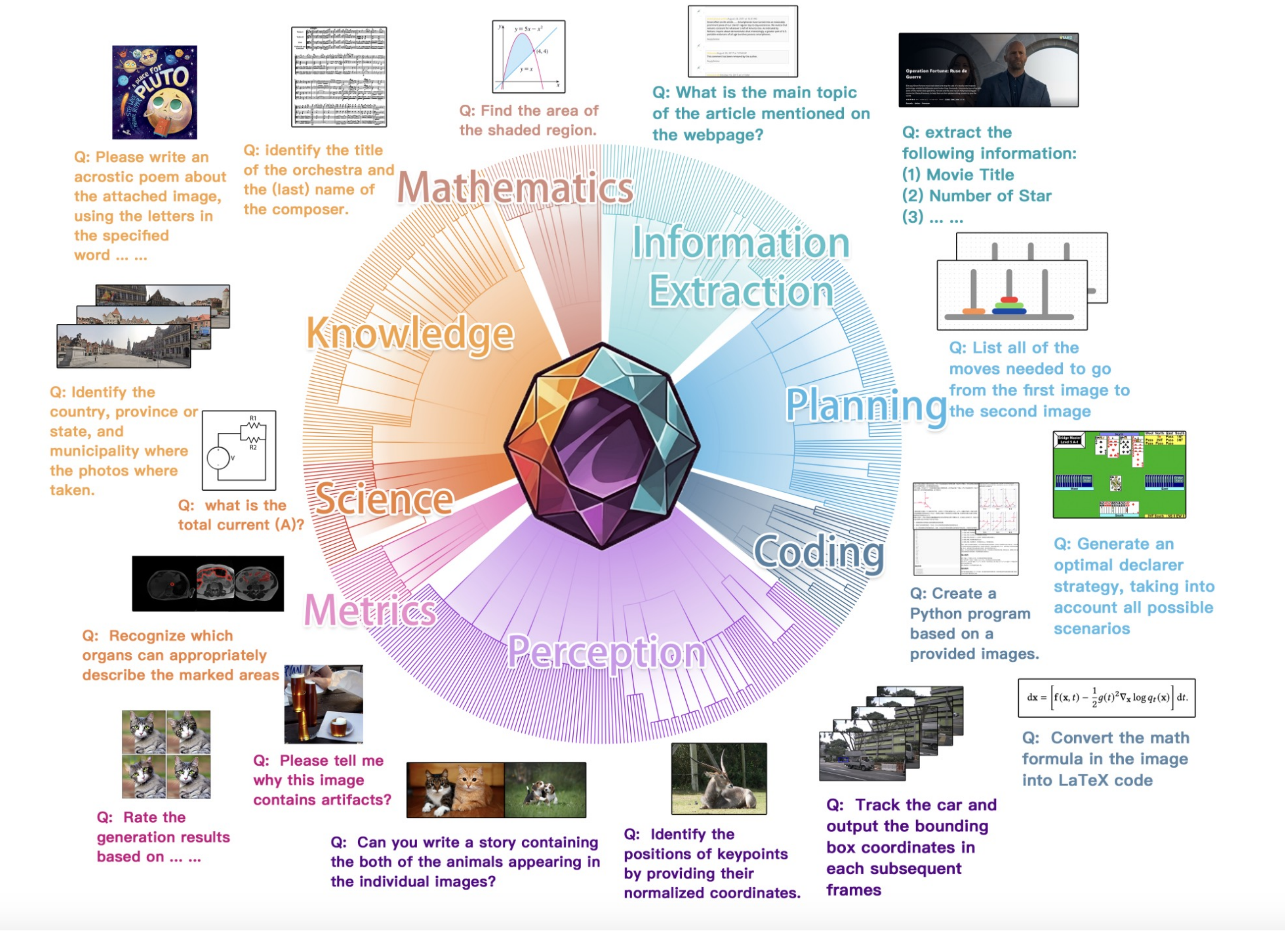
Correlation between average benchmark scores and models' loss

Hoffmann et al., Training Compute-Optimal Large Language Models, NeurIPS'2022.
Huang et al., Compression Represents Intelligence Linearly, COLM'2024.

Towards General Artificial Intelligence

- The scaled-up language models acquire a wide spectrum of capabilities.





Wei et al., Emergent Abilities of Large Language Models, TMLR'2022.
Chen et al., MEGA-Bench: Scaling Multimodal Evaluation to over 500 Real-World Tasks, arXiv'2024.

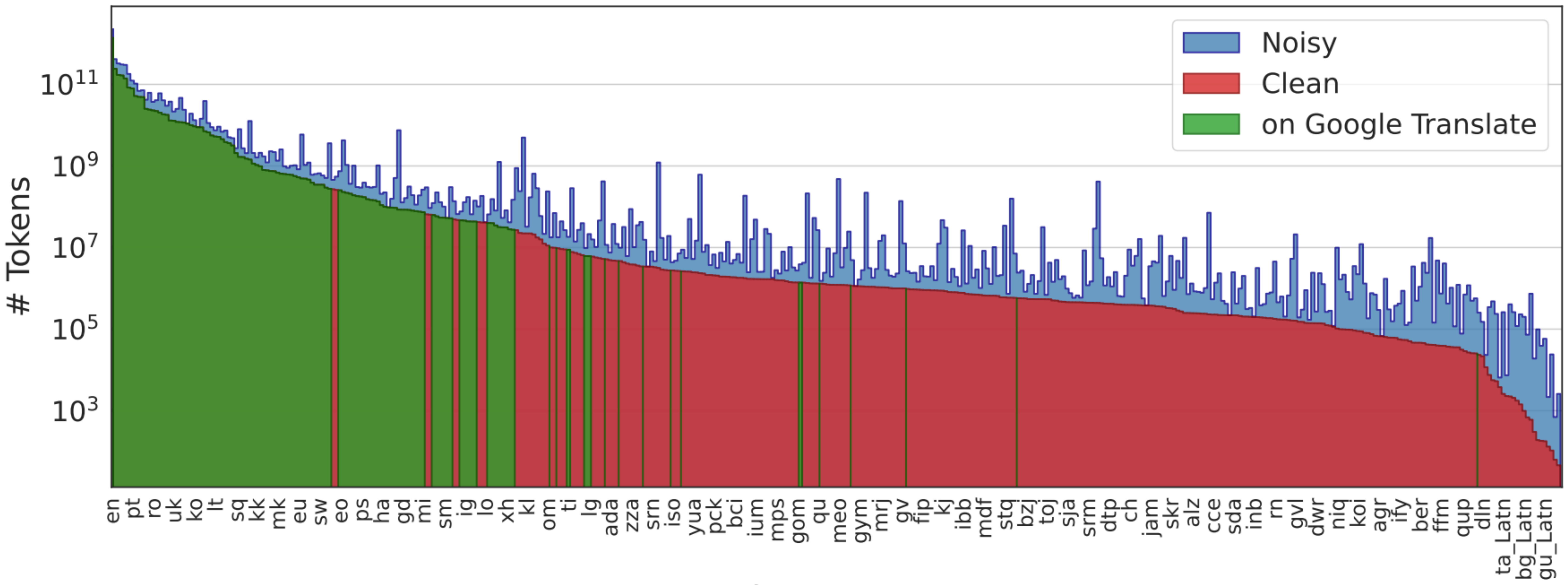
Unbalanced Data Distribution

- Scaling up language model makes data imbalance issues severe.
  - relatively easy to collect English corpus
  - hard to collect non-English corpus

LLaMA2

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

MADLAD-400



Touvron et al., LLaMA2: Open Foundation and Fine-Tuned Chat Models, arXiv'2023.
Kudugunta et al., MADLAD-400: A Multilingual And Document-Level Large Audited Dataset, NeurIPS'2023.

Unbalanced Data Distribution

- Scaling up language model makes data imbalance issues severe.
  - hard to collect non-English corpus
  - hard to filter non-English pages (②)
  - hard to identify page languages (③)
  - hard to filter non-English sentences (④)

heavily rely on native speakers' observations and annotations

Kudugunta et al., MADLAD-400: A Multilingual And Document-Level Large Audited Dataset, NeurIPS'2023.

Unbalanced Data Distribution

- Most of post-training datasets, including instruction and preference data also focus on English.

Datasets	Sourced from	# Instances	$\bar{N}_{rounds}$	$\bar{L}_{prompt}$	$\bar{L}_{completion}$
SuperNI [48]	NLP datasets + Human-written Instructions	96,913	1.0	291.1	38.7
CoT [50]	NLP datasets + Human-written CoTs	100,000	1.0	266.0	53.2
Flan V2 [31]	NLP datasets + Human-written Instructions	100,000	1.0	355.7	31.2
Dolly [12]	Human-written from scratch	15,011	1.0	118.1	91.3
Open Assistant 1 [26]	Human-written from scratch	34,795	1.6	34.8	212.5
Self-instruct [47]	Generated w/ vanilla GPT3 LM	82,439	1.0	41.5	29.3
Unnatural Instructions [23]	Generated w/ Davinci-002	68,478	1.0	107.8	23.6
Alpaca [43]	Generated w/ Davinci-003	52,002	1.0	27.8	64.6
Code-Alpaca [6]	Generated w/ Davinci-003	20,022	1.0	35.6	67.8
GPT4-Alpaca [36]	Generated w/ Davinci-003 + GPT4	52,002	1.0	28.0	161.8
Baize [52]	Generated w/ ChatGPT	210,311	3.1	17.6	52.8
ShareGPT[3]	User prompts + outputs from various models	168,864	3.2	71.0	357.8

More than 7,000 languages[1] are spoken around the world today, with a considerable number facing the challenges of being low-resourced, under-represented, or disappearing [Maxwell & Hughes, 2006; Simons, 2019; Moran & Chiarcos, 2020; Secretariat, 2022; Gao & Liu, 2023; Ilhomovna & Yuldasheva, 2023; Marivate et al., 2020]. In contrast, the most widely used datasets and breakthroughs in NLP have coalesced around a few data-rich languages [Longpre et al., 2023b; Taori et al., 2023; Chung et al., 2022; Fan et al., 2021; Dodge et al., 2021; Lucy et al., 2024]. IFT datasets are no exception; the creation of these datasets has almost entirely focused on English. Furthermore, the vast majority of the creators of these works originate from a few countries [Longpre et al., 2023b; Zhang et al., 2022].

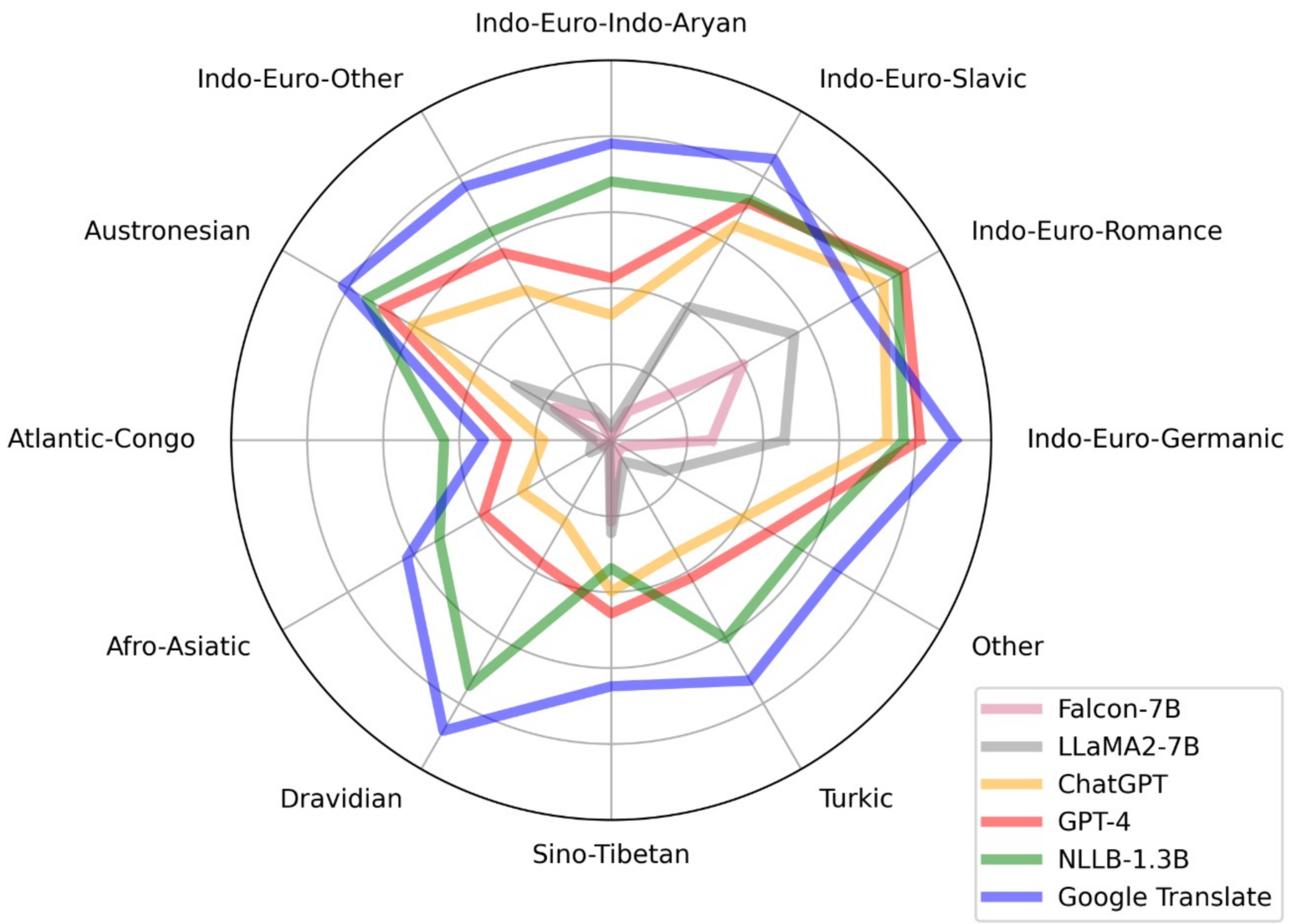
Wang et al., How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources, NeurIPS'2023.
Singh et al., Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning, arXiv'2024.

Tutorial Roadmap

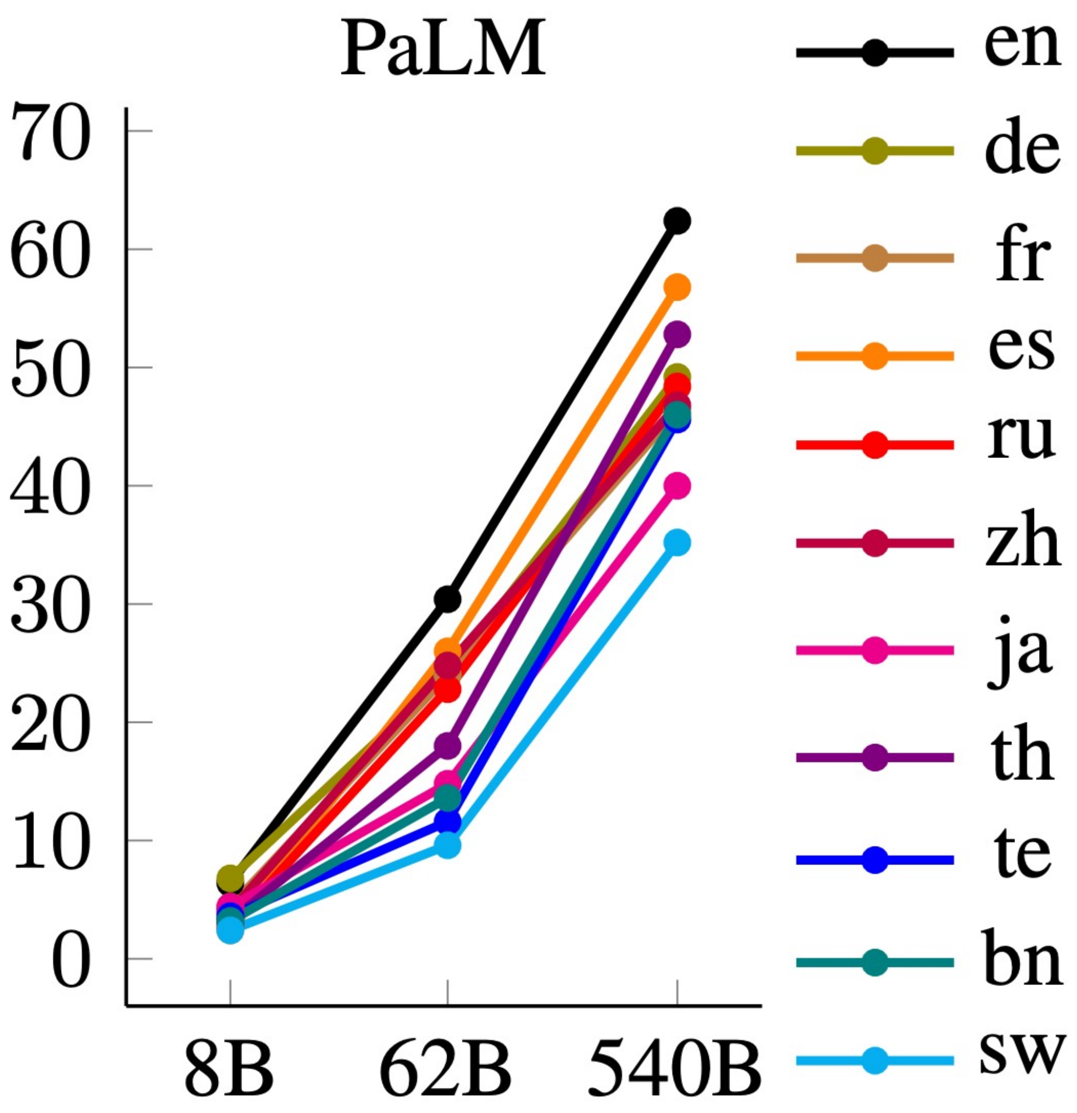
- Chapter I: Background
- Chapter II: Observations and Analyses
- Chapter III: Enhancing LLM for More Languages
- Chapter IV: Aligning Non-English to English
- Chapter V: Future Challenges

Scaling up does not solve Multilingualism

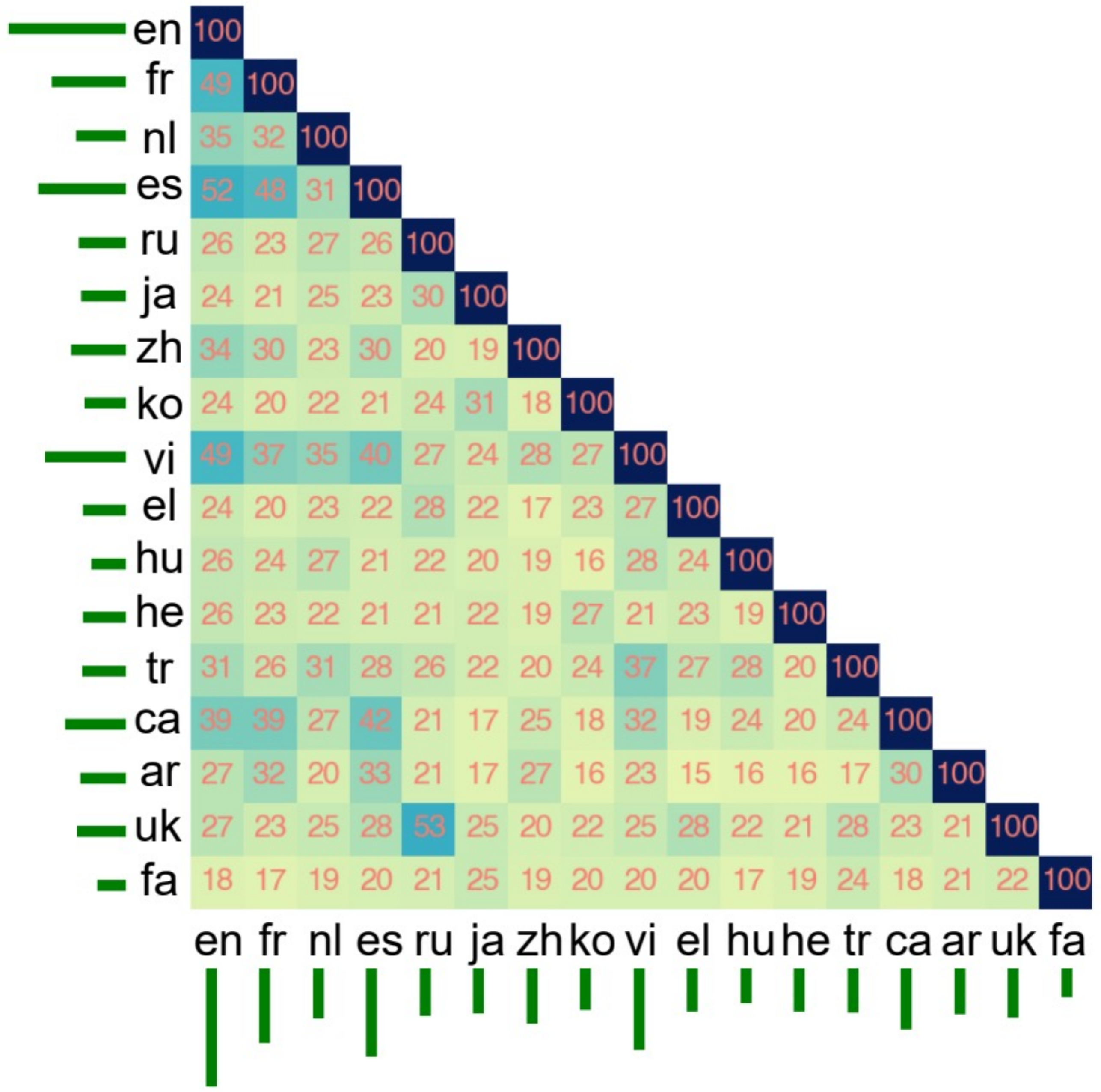
- LLM performs much worse in translation, reasoning, factual consistency, etc., for languages that are dissimilar to English, such as Asian languages and African languages.



translation (Zhu et al.)



reasoning (Shi et al.)

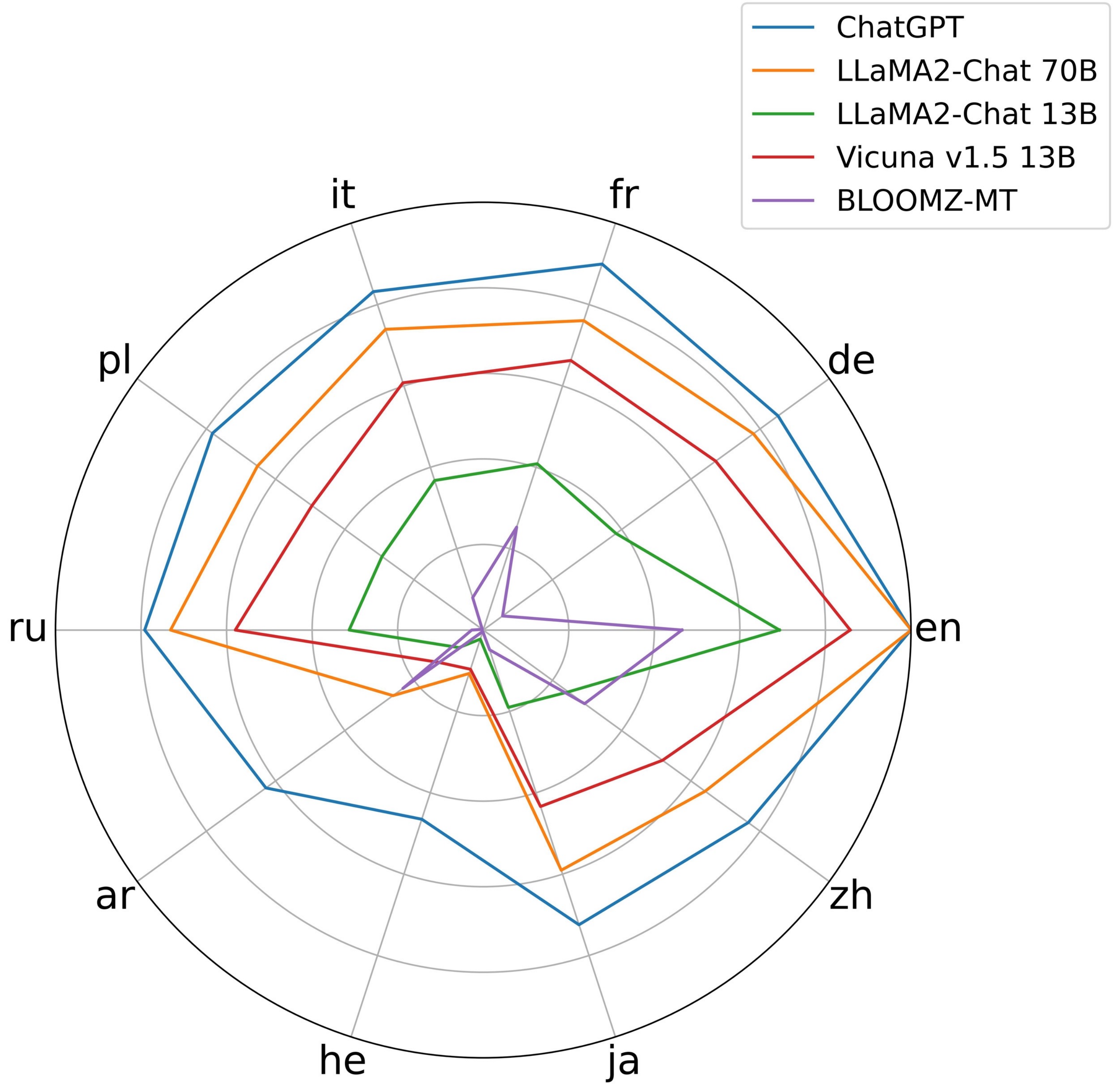


knowledge (Qi et al.)

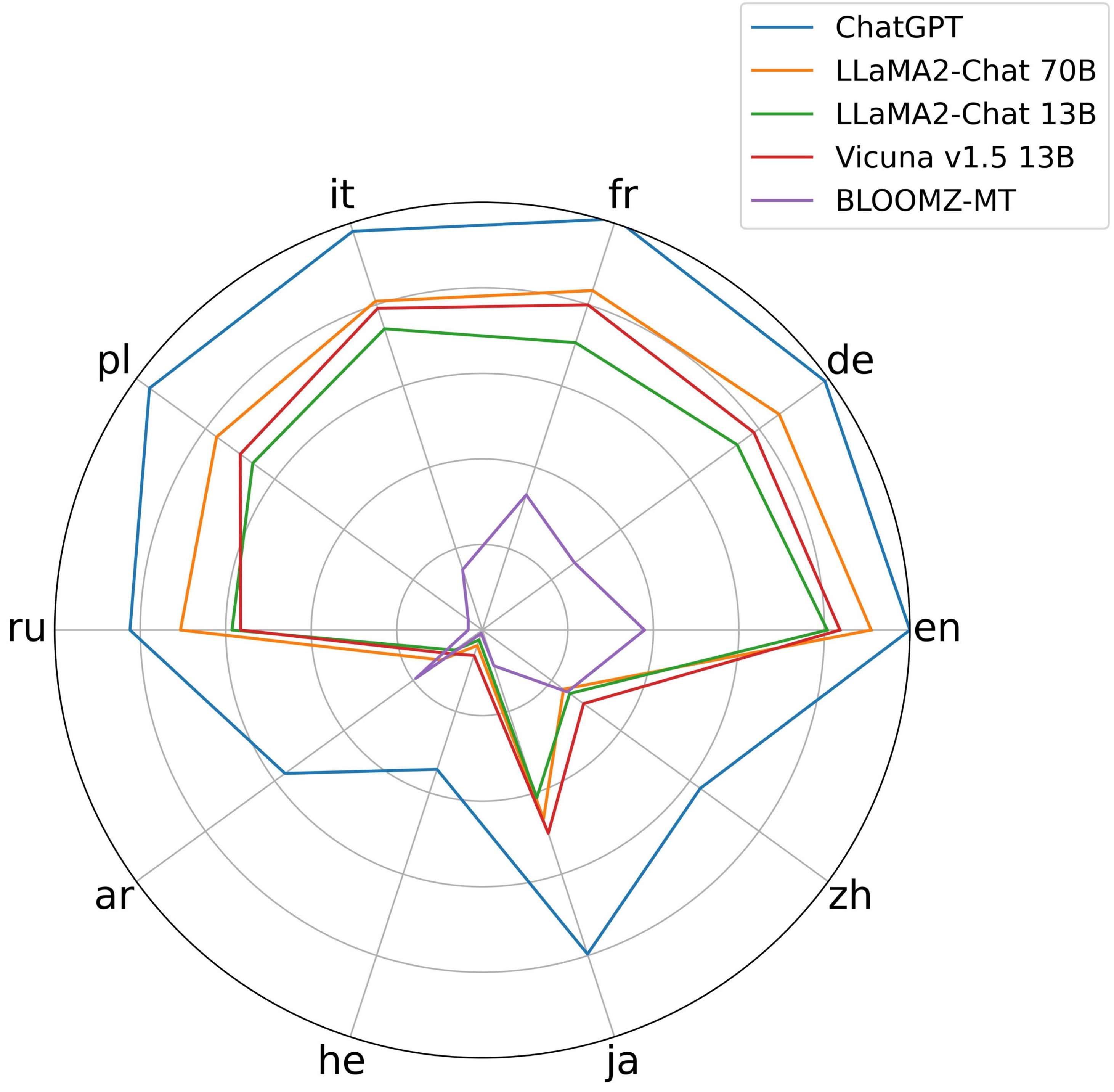
Shi et al., Language Models Are Multilingual Chain-Of-Thought Reasoners, ICLR'2023.
Qi et al., Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models, EMNLP'2023.
Zhu et al., Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis, Findings of NAACL'2024.

Knowledge related tasks

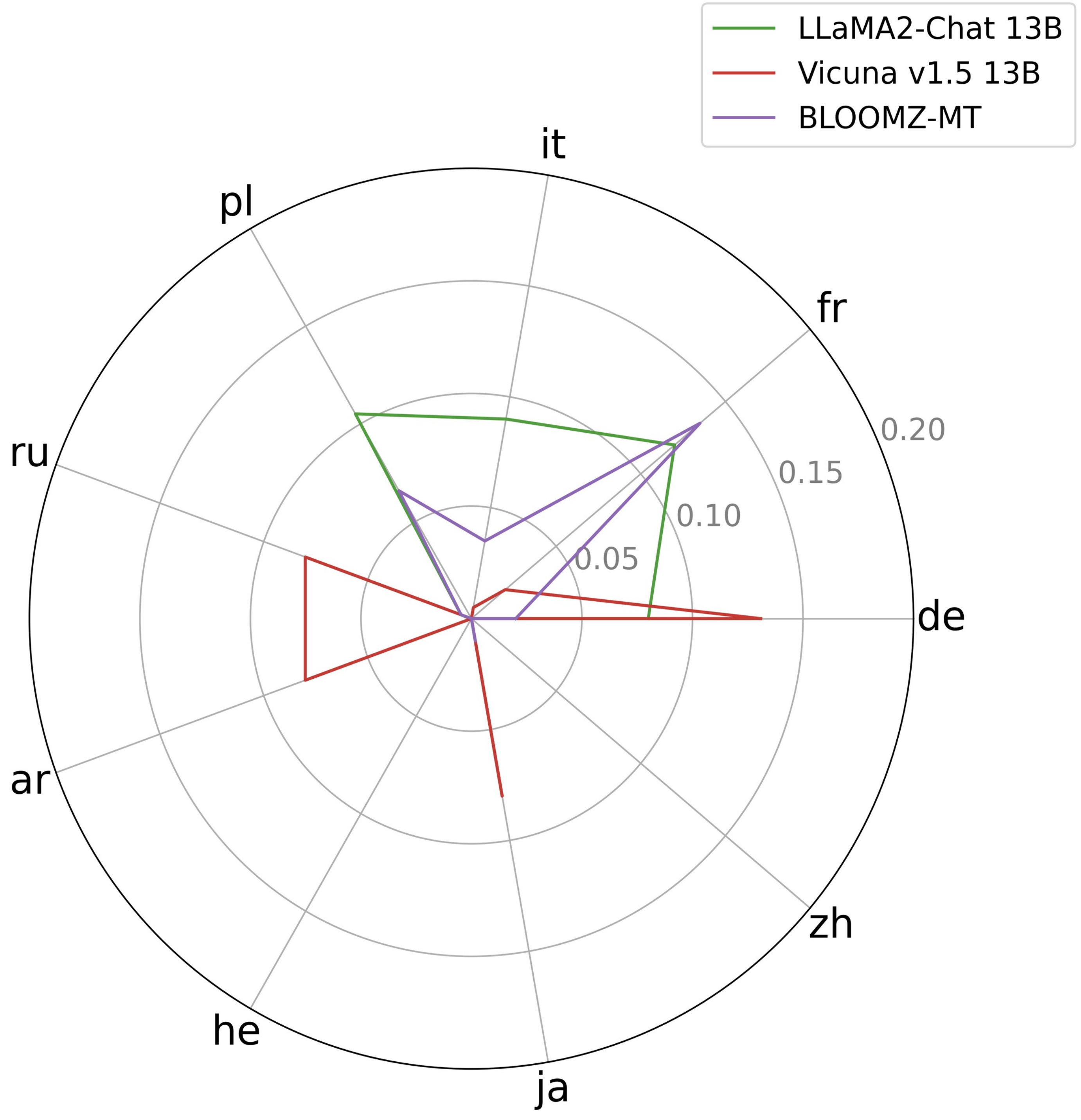
- LLMs perform diversely for different languages in knowledge related tasks.
  - especially poor when retrieving knowledge from other languages.



Common Sense QA



Factual QA

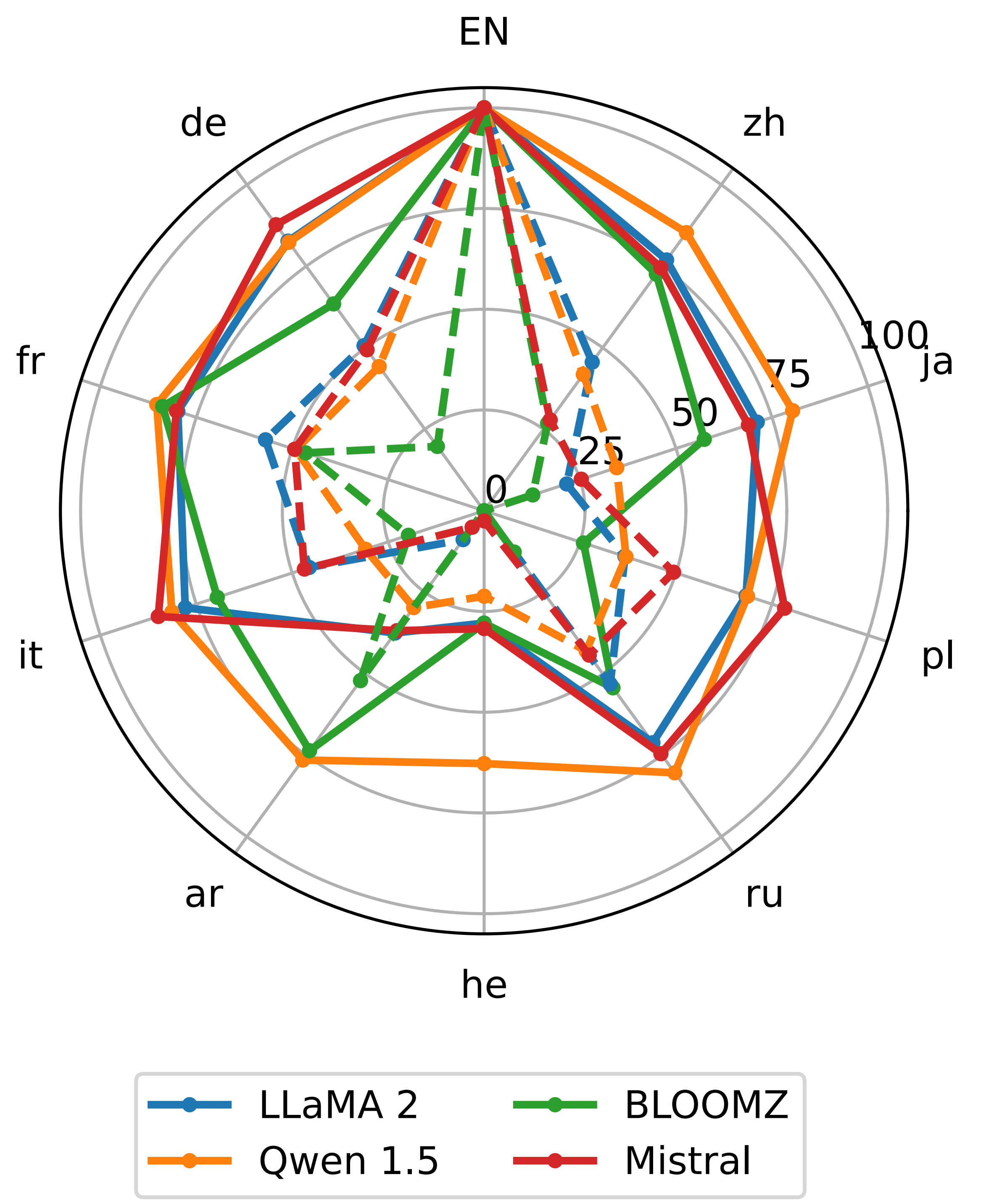


Cross-lingual QA

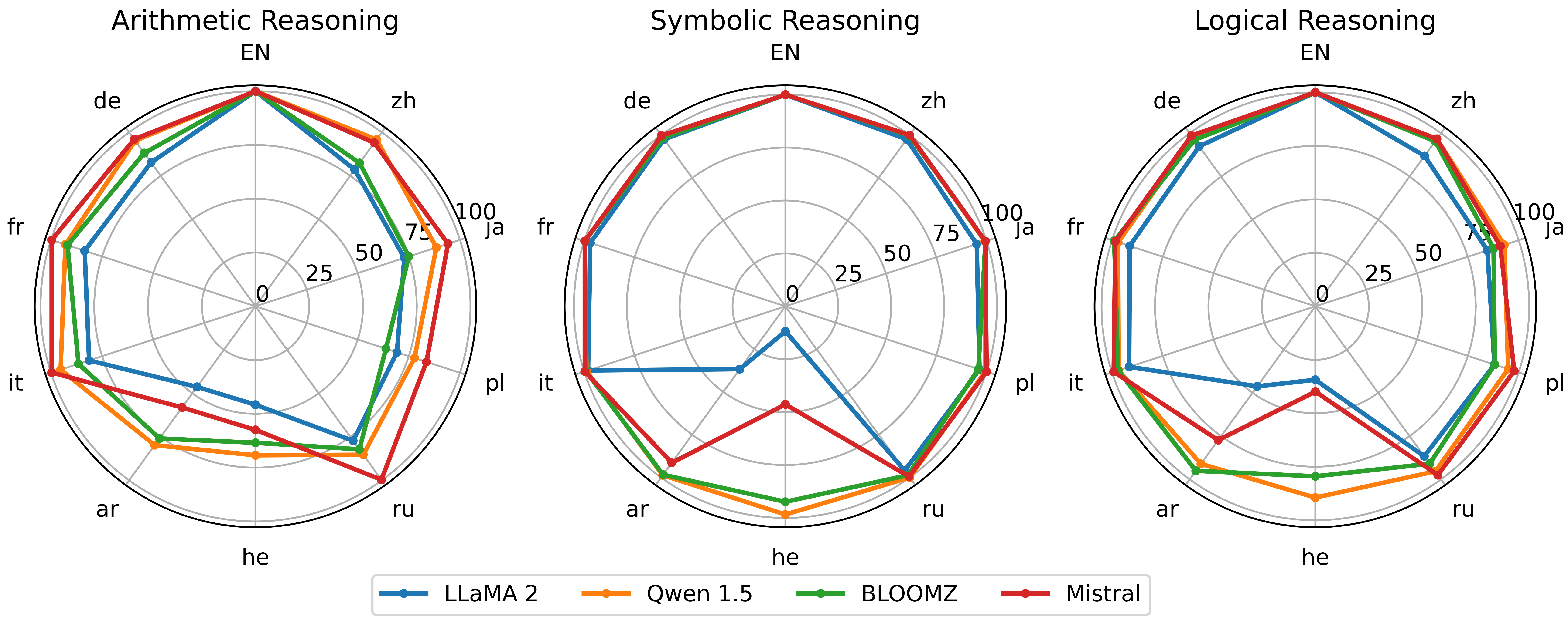
Gao et al., Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly, NAACL'2024.

Knowledge v.s. Cross-lingual Transfer

- Knowledge retrieving seems to be the key obstacle in cross-lingual transfer, i.e. performance in another language when trained in one.
- Knowledge-free tasks are better transferred to other languages.



Knowledge Related Tasks



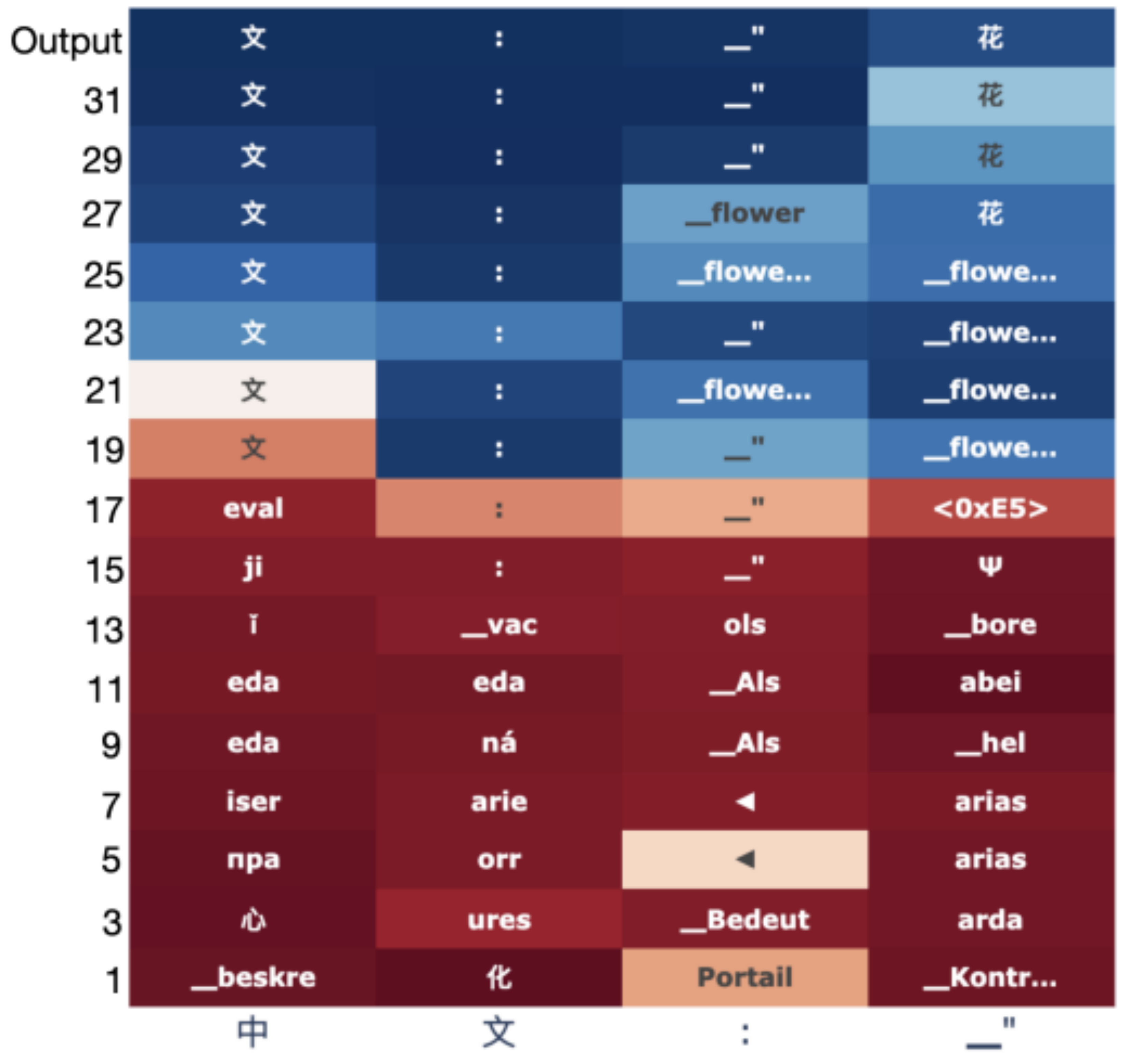
Knowledge-free Tasks

Hu et al., Large Language Models Are Cross-Lingual Knowledge-Free Reasoners, arxiv'2024.

Understanding LLM's Multilingual Working Pattern

- Wendler et al. discover that when LLMs perform multilingual tasks, they show a three-phase working pattern.

- by observing layer-wise logit lens:
  - phase 1: grounding to non-sense tokens
  - phase 2: grounding to English tokens
  - phase 3: grounding to non-English tokens



Phase 3

Phase 2

Phase 1

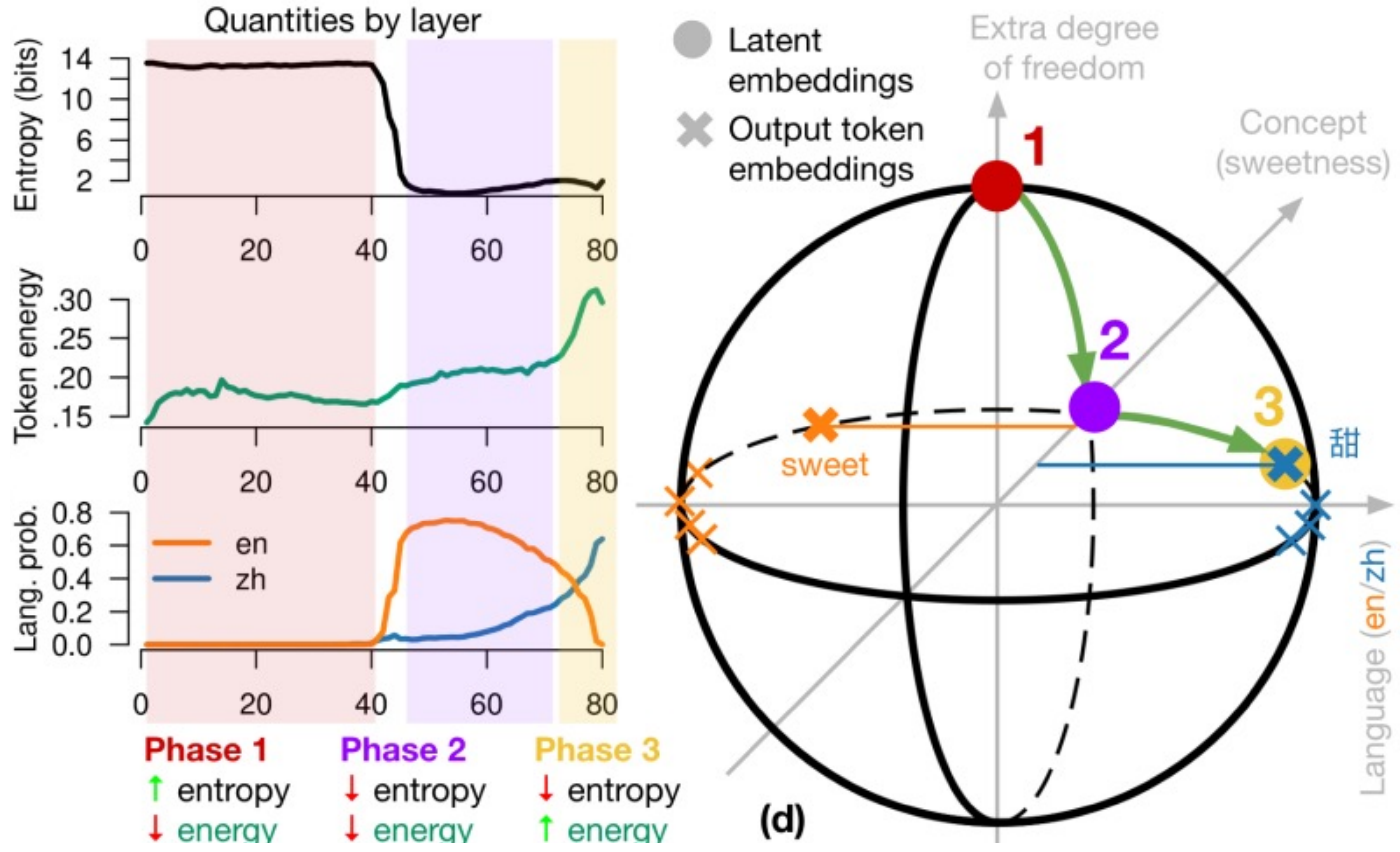
Français: "vertu" - 中文: "德"
Français: "siège" - 中文: "座"
Français: "neige" - 中文: "雪"
Français: "montagne" - 中文: "山"
Français: "fleur" - 中文: "

Wendler et al., Do Llamas Work in English? On the Latent Language of Multilingual Transformers, ACL'2024.

Understanding LLM's Multilingual Working Pattern

- layer-wise qualitative fine-grained observation
  - entropy: whether latents are orthogonal to output token space
  - token energy: how much of the latent is relevant for predicting the next token
  - language probability: the prob that the grounded token belongs to a specific language

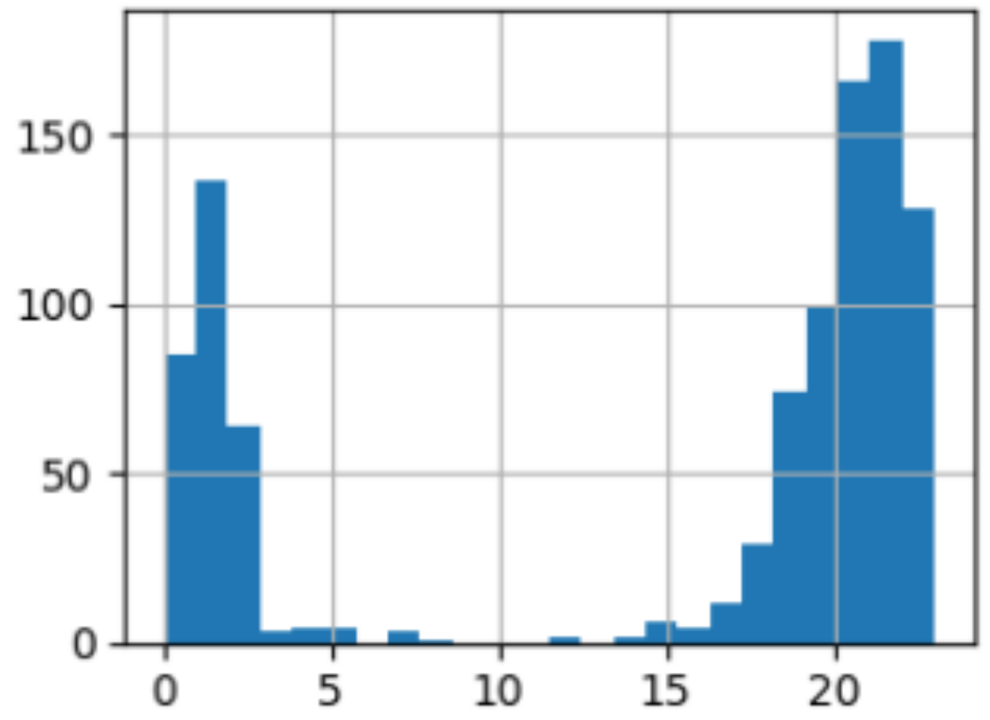
- proposed three-phase working pattern
  - phase 1: context understanding
  - phase 2: concept processing
  - phase 3: token generation



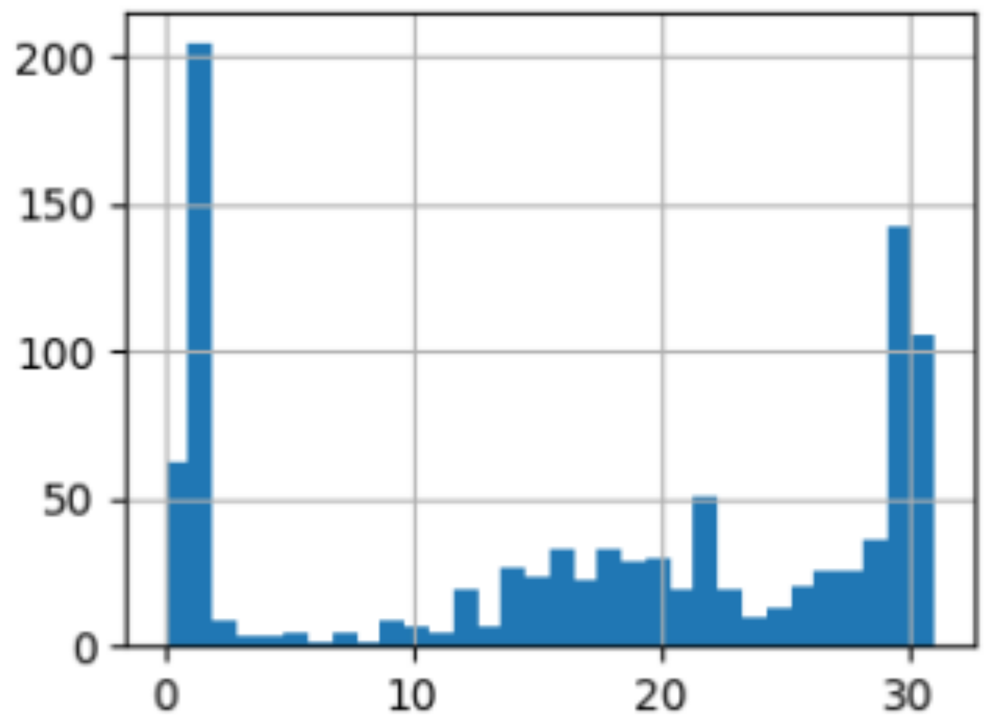
Wendler et al., Do Llamas Work in English? On the Latent Language of Multilingual Transformers, ACL'2024.

Neuro-level Analysis

- Language specific neurons are mainly distributed in the models' first and last few layers (Bhattacharya & Bojar., Kojima et al., Tang et al.).
- The models can control language in text generation by intervening with language-specific neurons (Kojima et al., Tang et al.).



BLOOM-1.7B (Fr)



LLaMA2-13B (Zh)

Without any intervention	Machu Picchu consist of three main structures, namely Intihuatana, the Temple of the Sun, and the Room of the Three Windows.
Intervention in German neurons	Machu Picchu besteht aus drei Hauptstrukturen, nämlich Intihuatana, der Tempel der Sonne und die Zimmer mit drei Fenstern.
Intervention in Chinese neurons	秘鲁的马腾岭有三个主要的建筑，即祭坛、圣殿和三窗房。

	"Translate a sentence from English to a target language."					
	Accuracy			BLEU		
de	0.0	→	**62.0**	2.8	→	**16.5**
es	5.0	→	**78.0**	4.0	→	**16.5**
ja	0.0	→	**55.0**	0.3	→	**9.2**
fr	0.0	→	**58.0**	3.4	→	**21.3**
zh	1.0	→	**79.0**	1.2	→	**12.7**

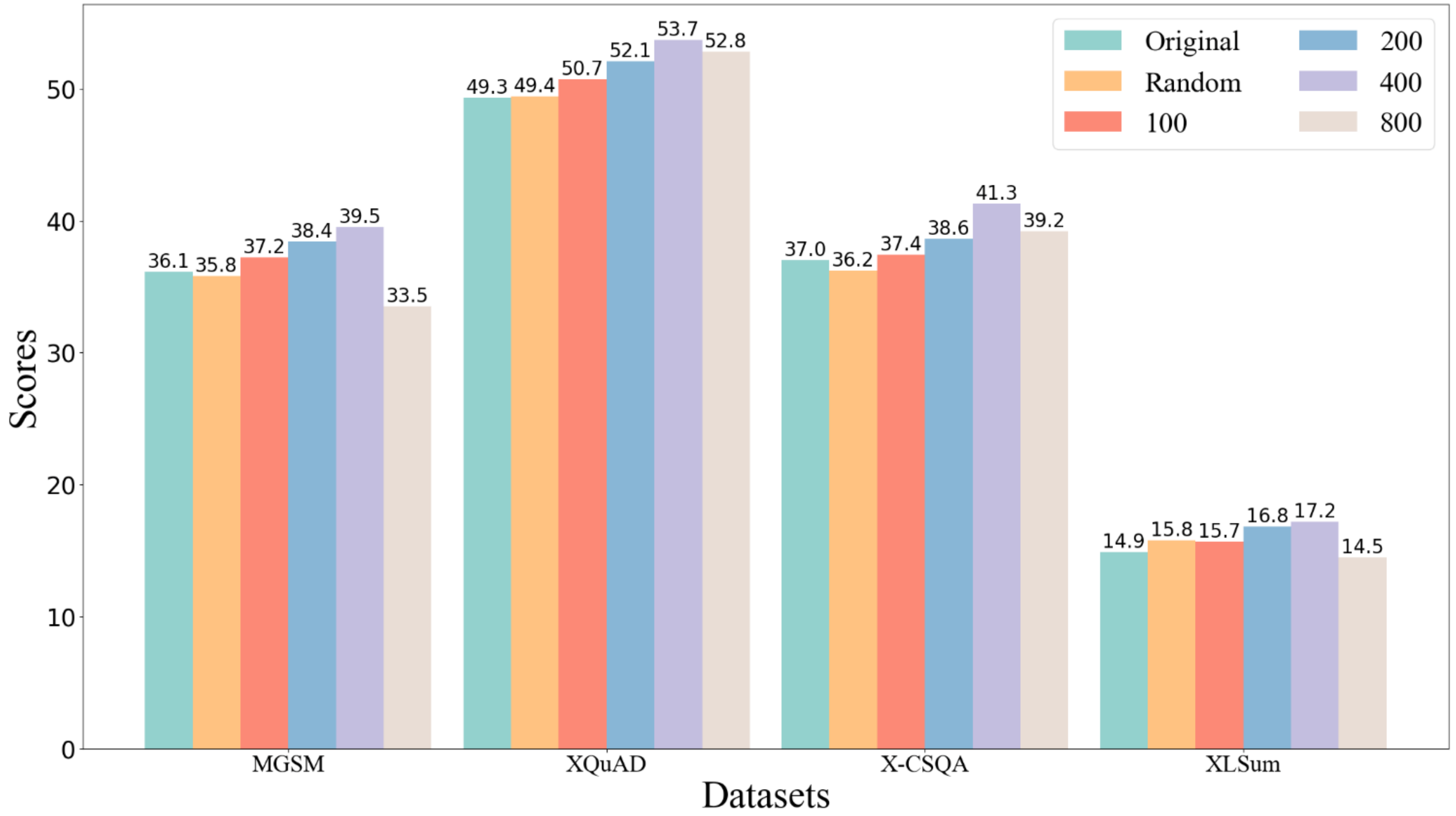
Bhattacharya & Bojar, Unveiling Multilinguality in Transformer Models: Exploring Language Specificity in Feed-Forward Networks, arXiv'2023.
Kojima et al., On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons, NAACL.2024.
Tang et al., Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models, ACL'2024.

Neuro-level Analysis

- Disabling language-specific neurons harms performance, whereas deactivating random neurons does not.

Model	Method	Fr	Zh	Es	Ru	Avg.
Vicuna	Original	14.2	61.1	10.4	20.8	26.6
	Deactivate Random	14.1	61.6	10.4	20.8	26.7
	Deactivate Lang-Spec	**0.83**	**0.00**	**0.24**	**0.42**	**0.37**
Mistral	Original	15.2	56.4	10.6	21.0	25.8
	Deactivate Random	15.4	55.9	10.2	21.2	25.7
	Deactivate Lang-Spec	**0.21**	**0.39**	**0.15**	**0.07**	**0.21**

- Fine-tuning language-specific neurons boost performance, whereas fine-tuning random neurons does not.



Zhao et al., How do Large Language Models Handle Multilingualism? arXiv'2024.

Analysis of Internal Representation

- Same concept in different languages may activate the same internal pattern.

Feature #34M/31164353 **Golden Gate Bridge** feature example

The feature activates strongly on English descriptions and associated concepts

in the Presidio at the end (that's the⏎
huge park right next to the Golden Gate
bridge), perfect. But not all people

repainted, roughly, every dozen years."
"while across the country in san fran
cisco, the golden gate bridge was

it is a suspension bridge and has similar
coloring, it is often⏎> compared to the
Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋は、ア
メリカ西海岸のサンフランシスコ湾と太平洋が
接続するゴールデンゲート海

골든게이트 교 또는 금문교 는 미국 캘리포니아주
골든게이트 해협에 위치한 현수교이다. 골든게이
트 교는 캘리포니아주 샌프란시

мост золотые ворота — висячий мост
через пролив золотые ворота. он со
единяет город сан-фран

And on relevant images as well

Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet (https://transformer-circuits.pub/2024/scaling-monosemanticity/).

Take-away

- Observed:
  - diverse performance across languages.
  - knowledge retrieval/transfer affects the performance.
- Analysis:
  - solving tasks in other languages may have a multi-phase working pattern.
  - there are language specific and language agonistic neurons.
  - there are shared patterns for the same concept in different languages.
- The step further:
  - how do the cross-lingual effects/transfer happen?
  - from understanding to improving multilingualism.

Tutorial Roadmap

- Chapter I: Background
- Chapter II: Observations and Analyses
- Chapter III: Enhancing LLM for More Languages
- Chapter IV: Aligning Non-English to English
- Chapter V: Future Challenges

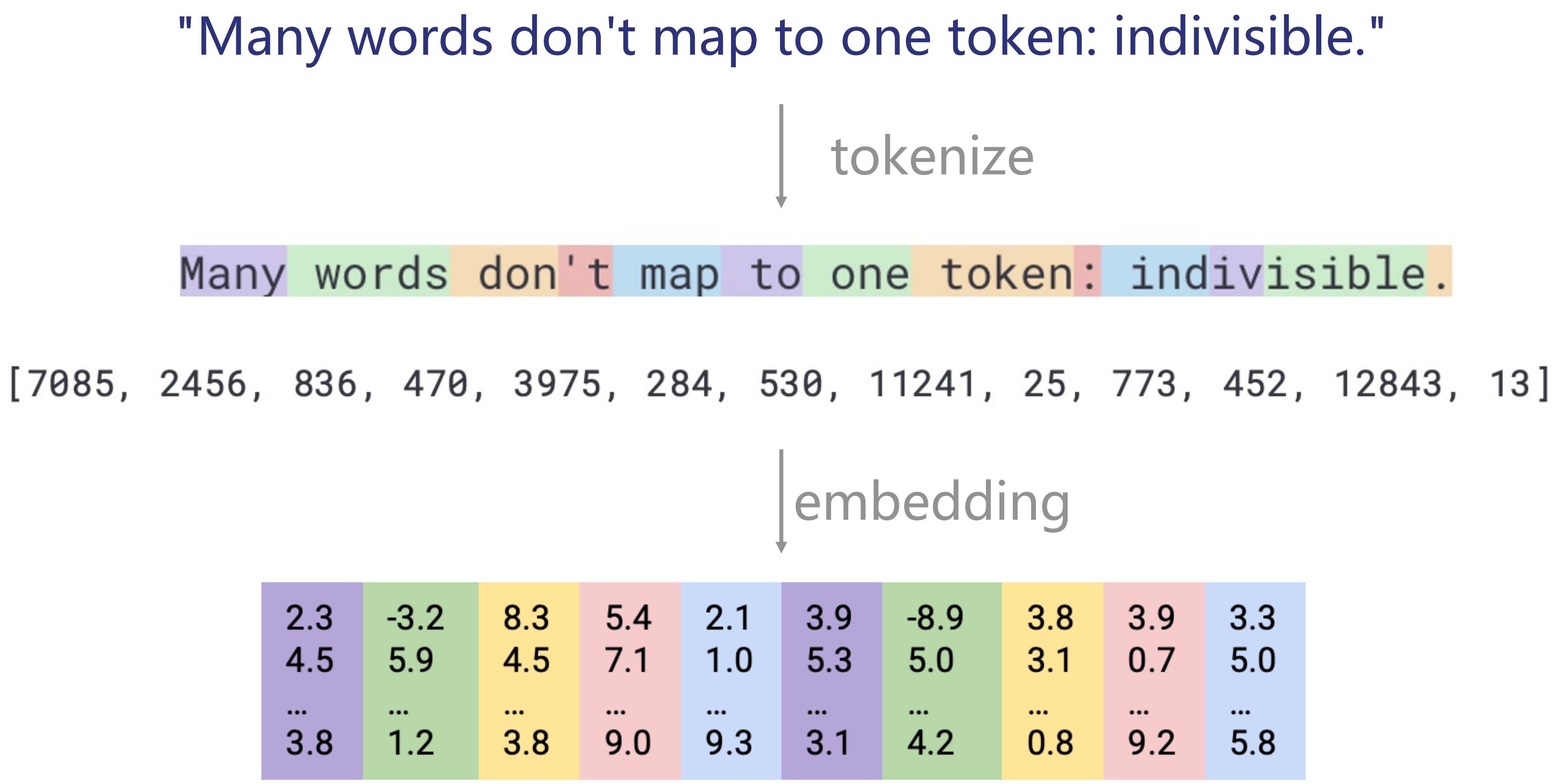
Improving LLM's Multilingual Performance

- Working pattern in three phrases
  - Phase 1: context understanding requires language understanding
  - Phase 2: concept processing requires knowledge retrieval, reasoning, ...
  - Phase 3: token generation requires language generation

- Knowing the Language(s) affects all three phases.
  - massive training with monolingual data
  - leveraging existing multilingual models
- Advance Abilities
  - multilingual post-training

Continual Pretraining Recipe

- tokenizer
  - fertility issue: it is expensive to process under-represented languages.
  - but vocabulary extension may have negative results (TowerLLM, LLaMaX).

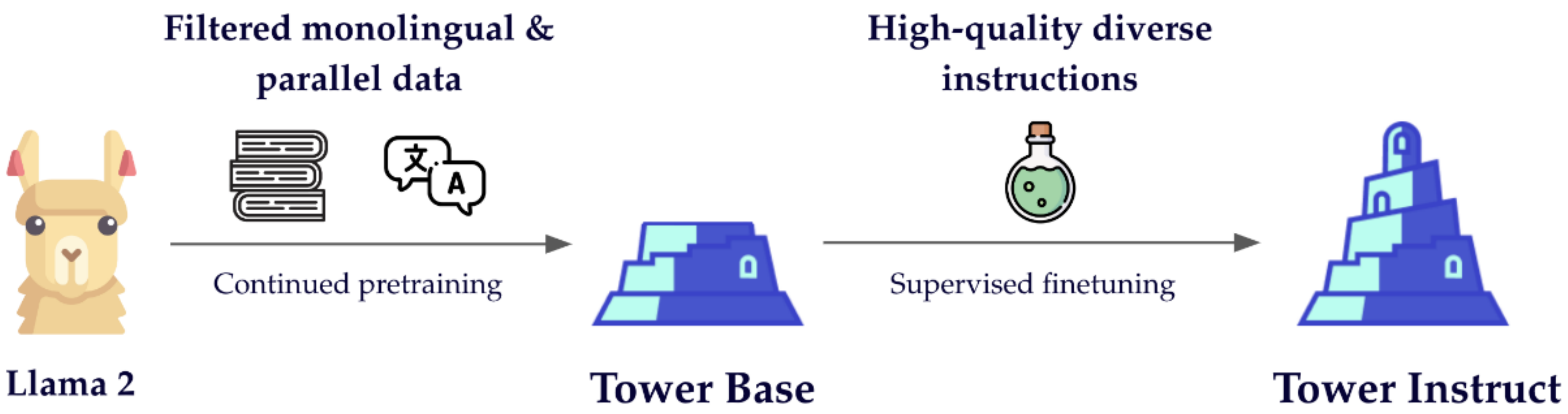


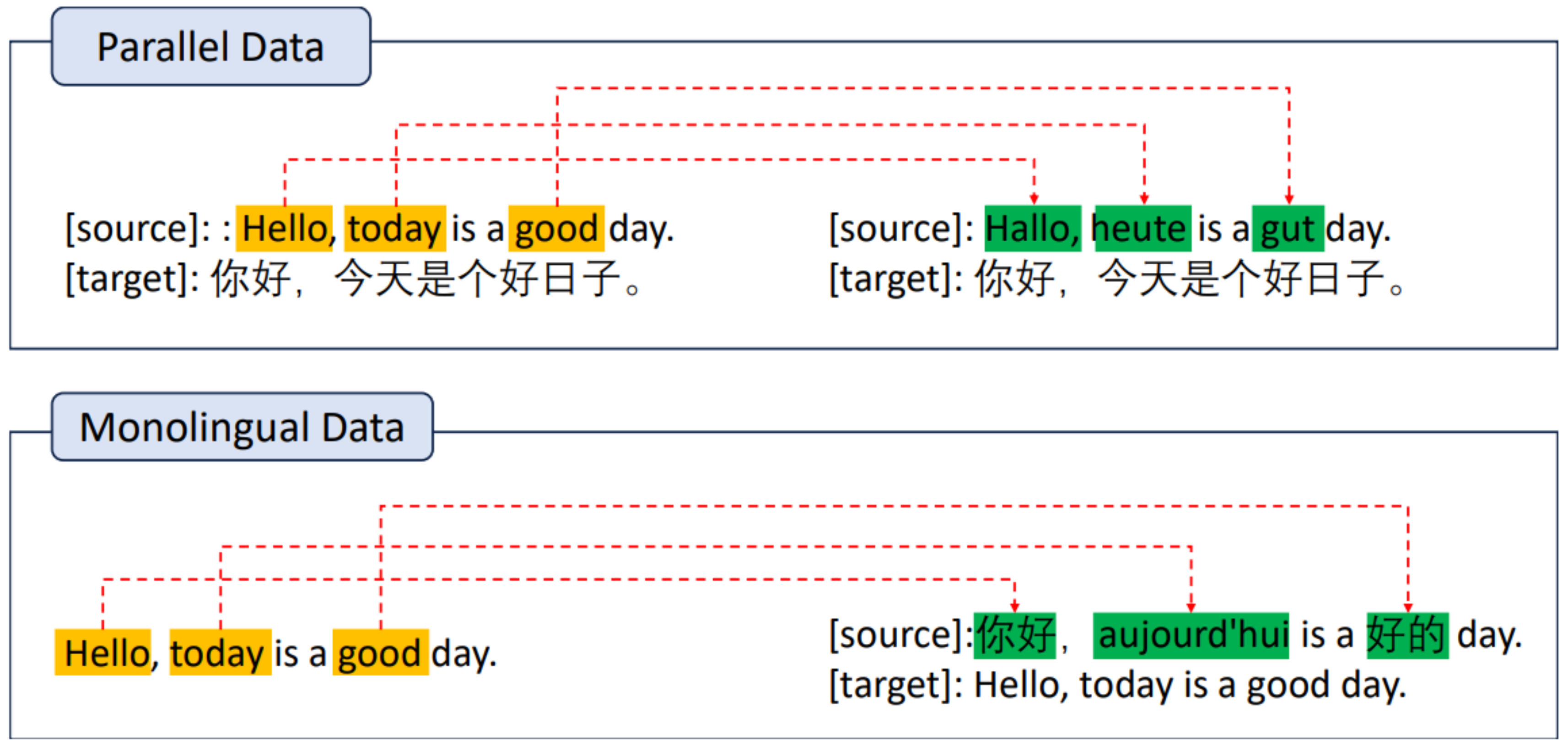
Language	ChatGPT's	Llama's
Vie	4.41	3.46
Zho	2.80	2.36
Tha	9.09	5.10
Ind	2.00	2.09
Khm	15.56	12.14
Lao	13.29	13.50
Msa	2.07	2.16
Mya	17.11	9.85
Tgl	2.28	2.22
Eng	1.00 (baseline)	1.19

Hyung Won Chung., Large Language Models (in 2023), invited talk@Seoul National University.
Alves et al., Tower: An Open Multilingual Large Language Model for Translation-Related Tasks, arXiv'2024
Lu et al., LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages, Findings of EMNLP'2024
Zhao et al., LLaMA Beyond English: An Empirical Study on Language Capability Transfer, arXiv'2024

Continual Pretraining Recipe

- data collection
  - multilingual monolingual resource, such as mC4, MADLAD-400, etc.
  - multilingual parallel resource, such as CC100, ParaCrawl, LegoMT, etc.
  - code-switched data augmentation

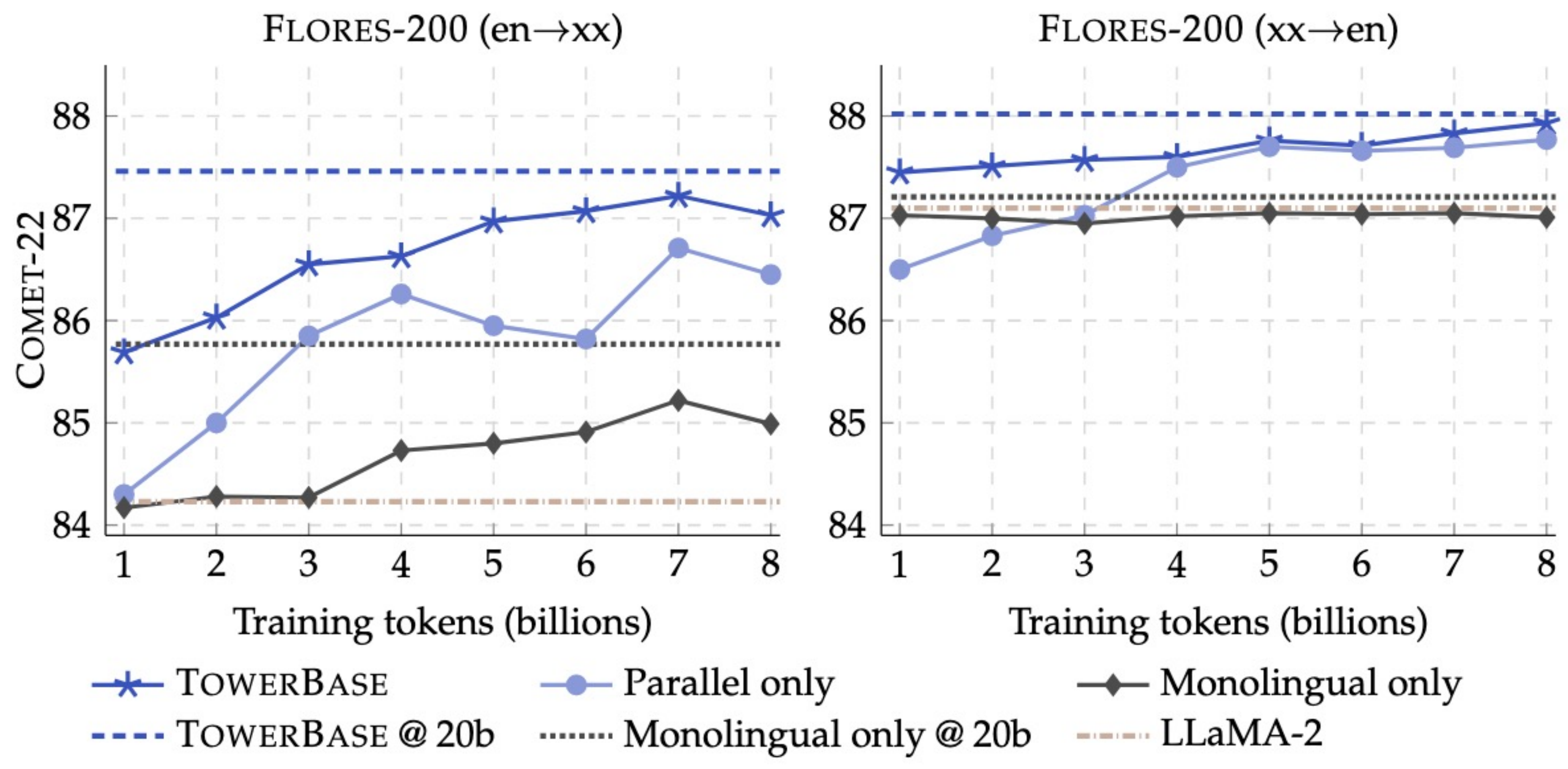




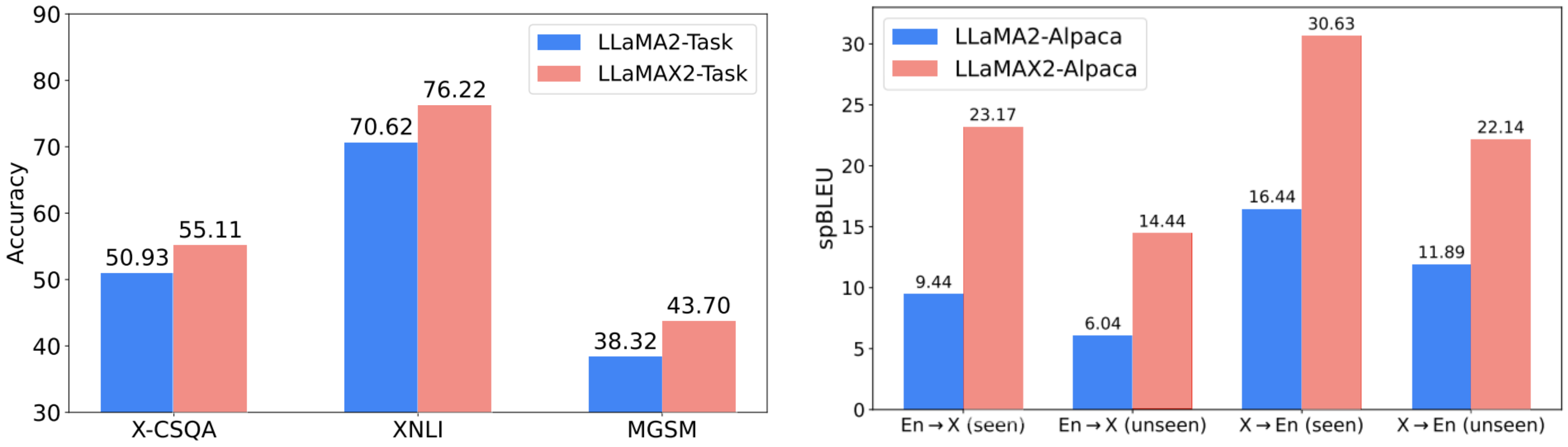
Xue et al., mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, NAACL'2021.
Kudugunta et al., MADLAD-400: A Multilingual And Document-Level Large Audited Dataset, arXiv'2023.
Banon et al., ParaCrawl: Web-Scale Acquisition of Parallel Corpora, ACL'2020.
Ji et al., EMMA-500: Enhancing Massively Multilingual Adaptation of Large Language Models, arXiv'2024.
Yuan et al., Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation, Findings of ACL'2023.

Continual Pretraining Recipe

- data mixture
  - incorporating English corpus avoids catastrophic forgetting
  - mixing monolingual and parallel data achieves the highest quality
  - generalize well even to unseen languages



TowerLLM (Alves et al.)

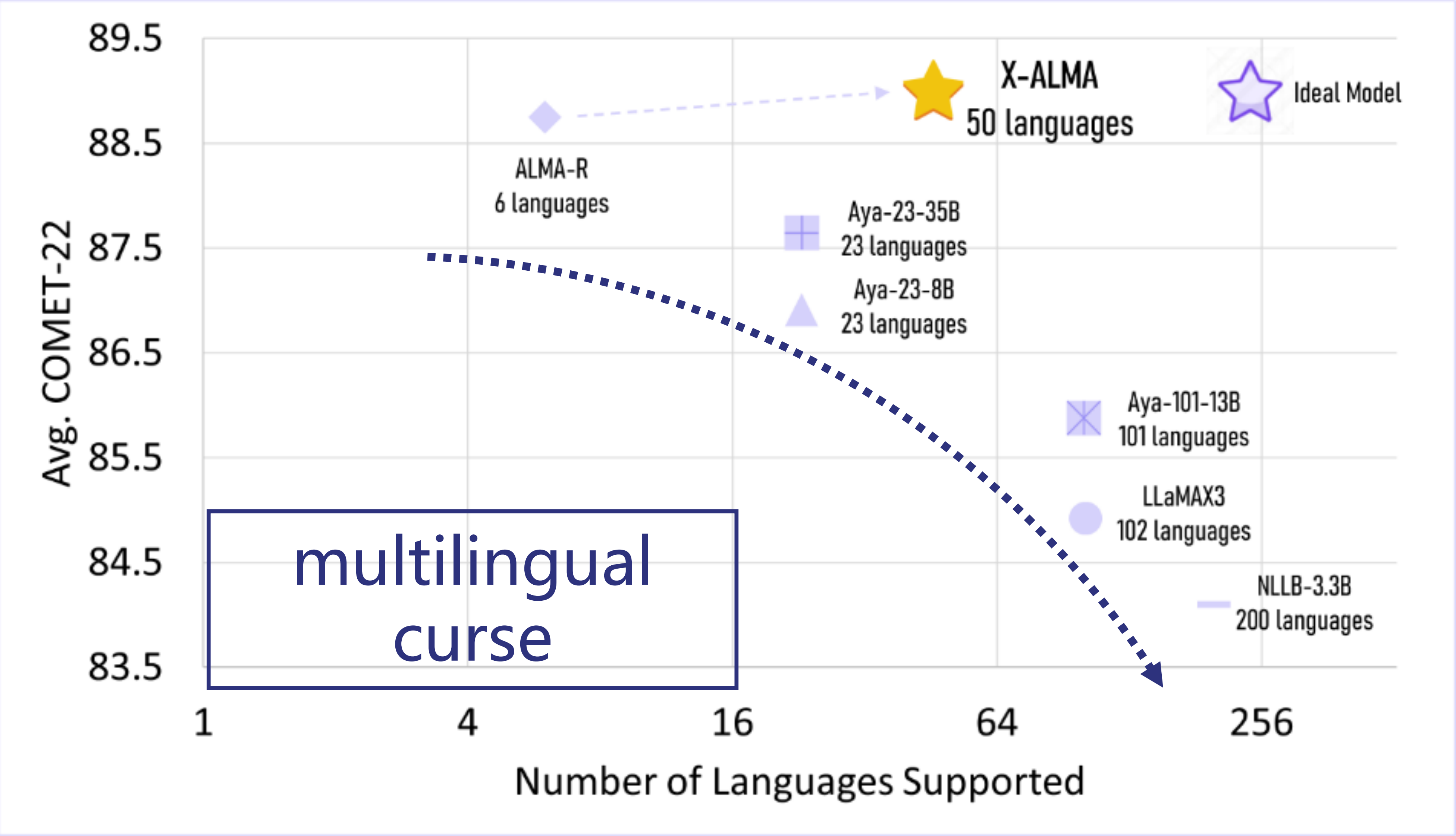


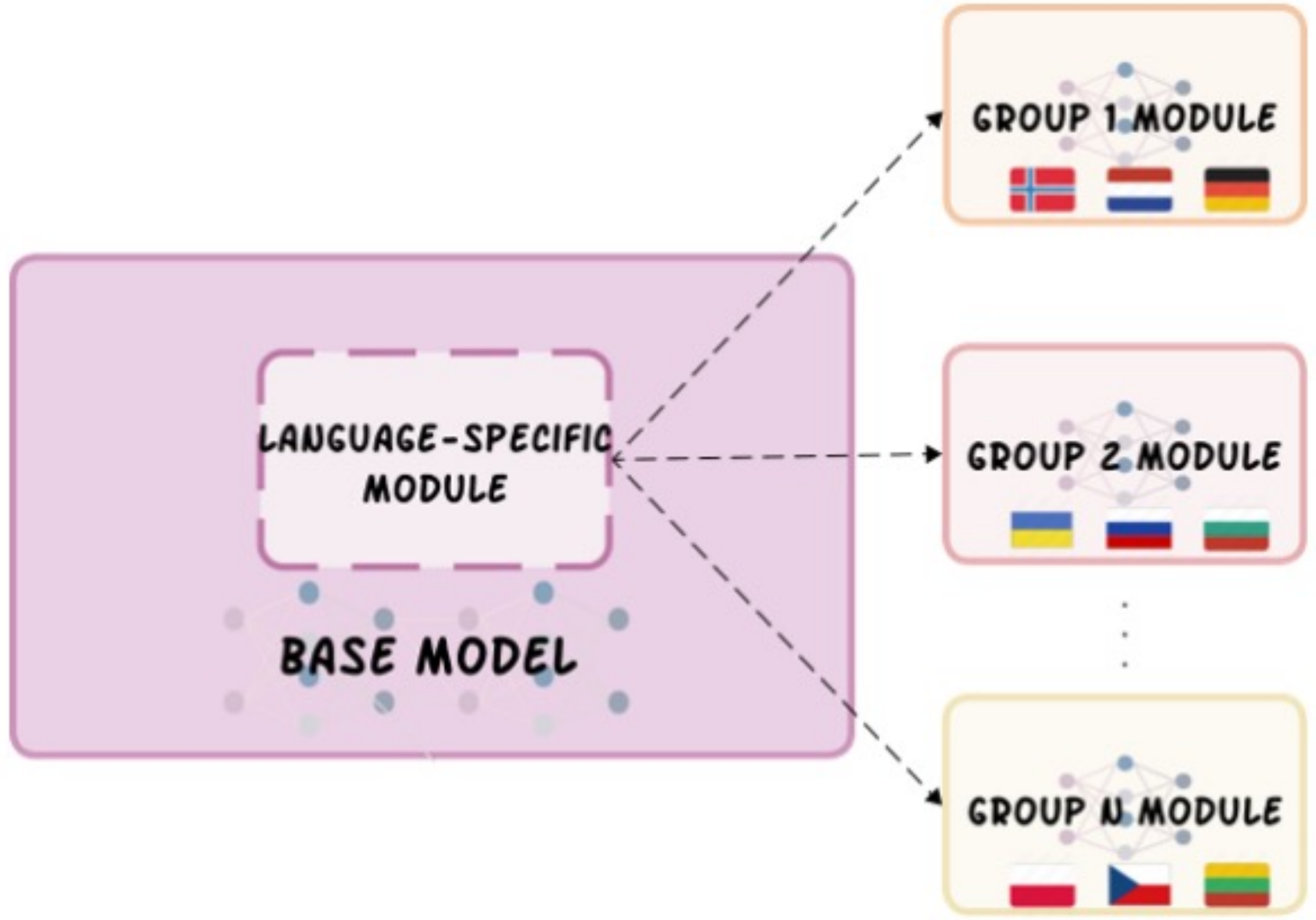
LLaMAX (Lu et al.)

Alves et al., Tower: An Open Multilingual Large Language Model for Translation-Related Tasks, arXiv'2024.
Lu et al., LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages, Findings of EMNLP'2024.

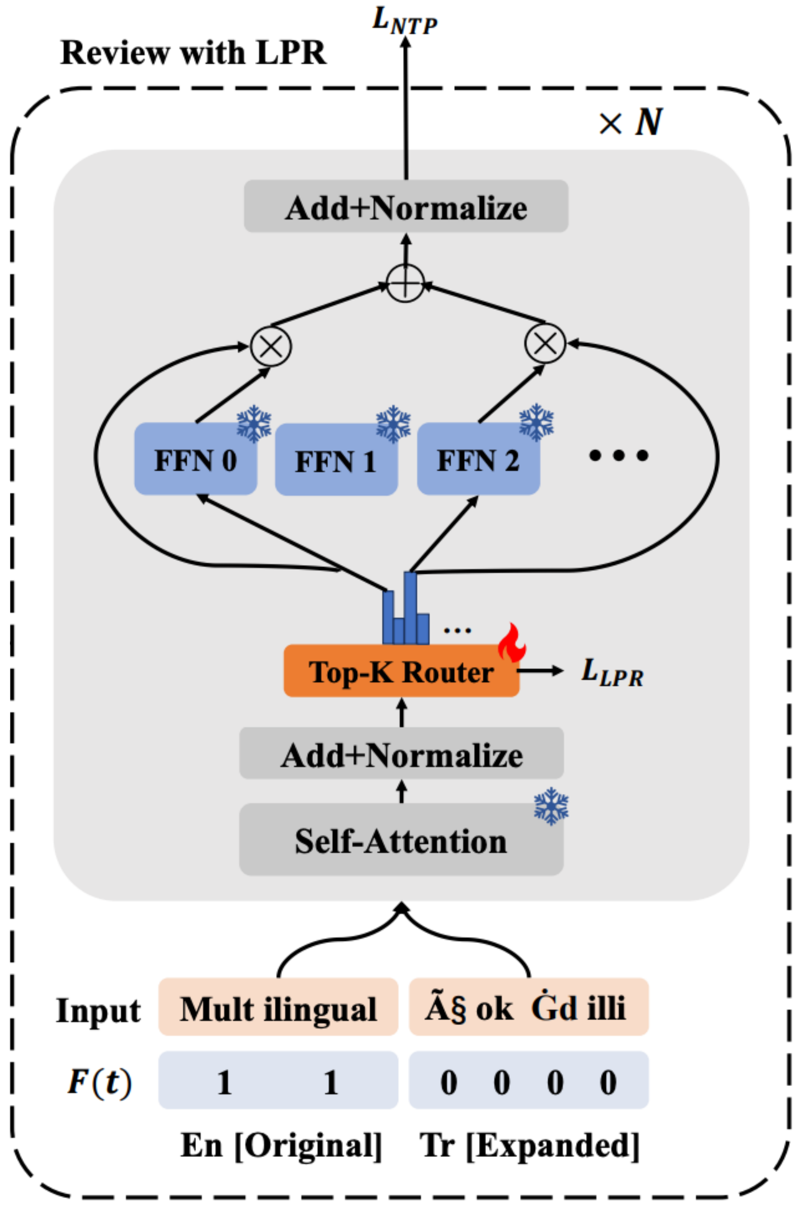
Continual Pretraining Recipe

- architecture: dense vs. sparse
  - all-in-one model may encounter "multilingual curse" (Conneau et al.).
  - recent efforts start to explore enhancing the base model with language-specific modules (Xu et al., Zhou et al.)





X-ALMA (Xu et al.)



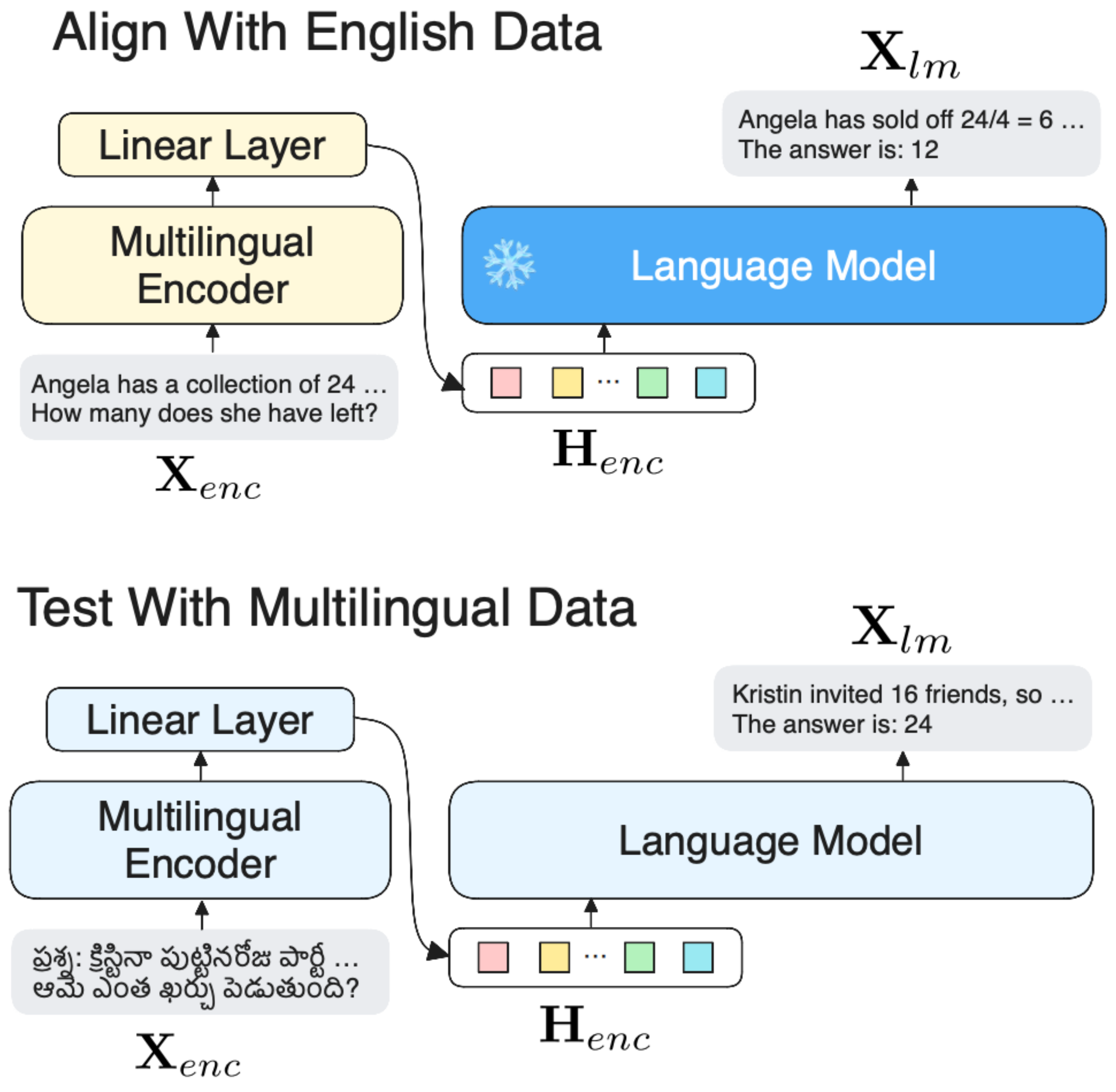
MoE-LPR (Zhou et al.)

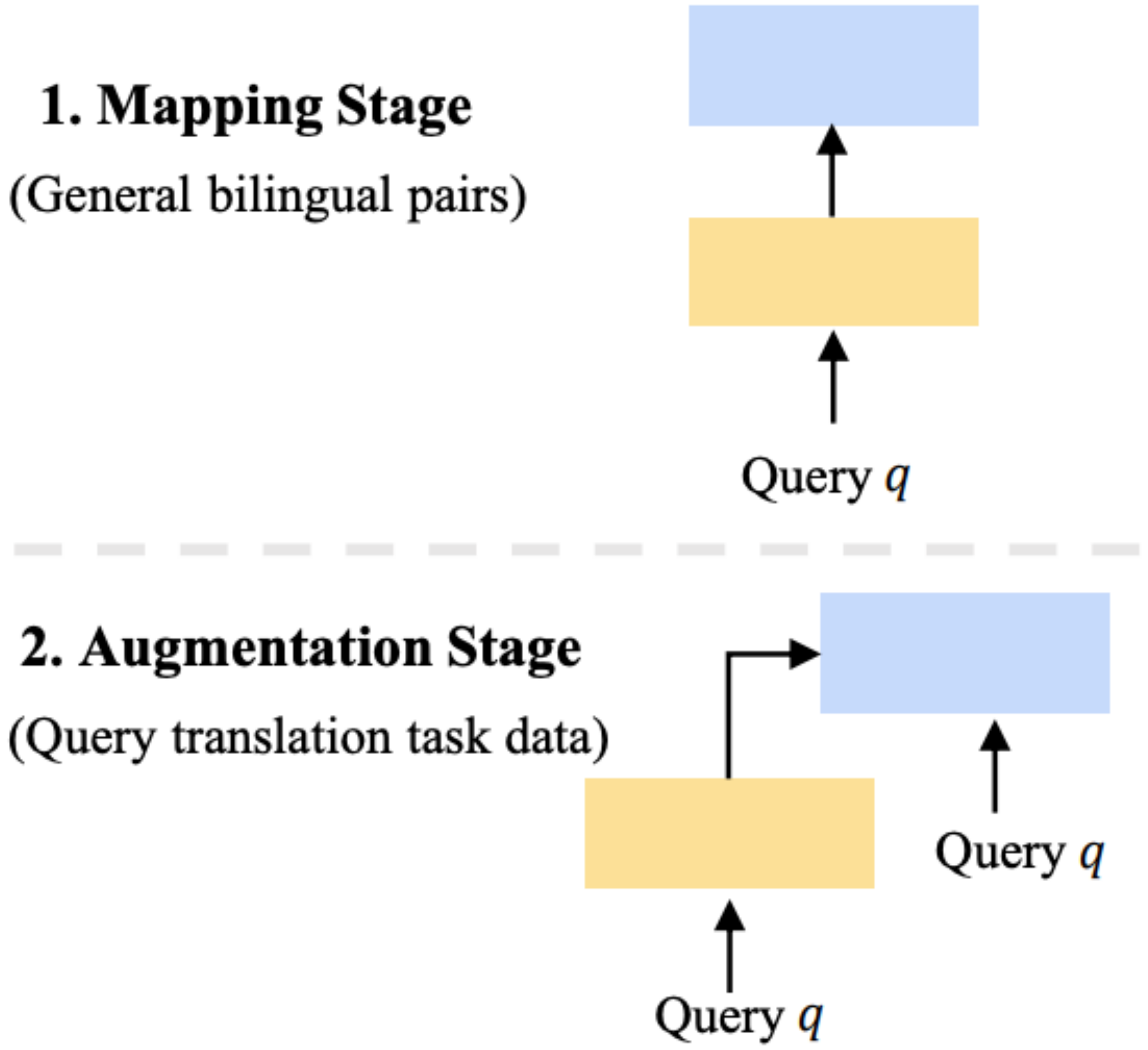
Conneau et al., Unsupervised cross-lingual representation learning at scale, ACL'2020.
Xu et al., X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale, arXiv'2024.
Zhou et al., MoE-LPR: Multilingual Extension of Large Language Models through Mixture-of-Experts with Language Priors Routing, arXiv'2024.

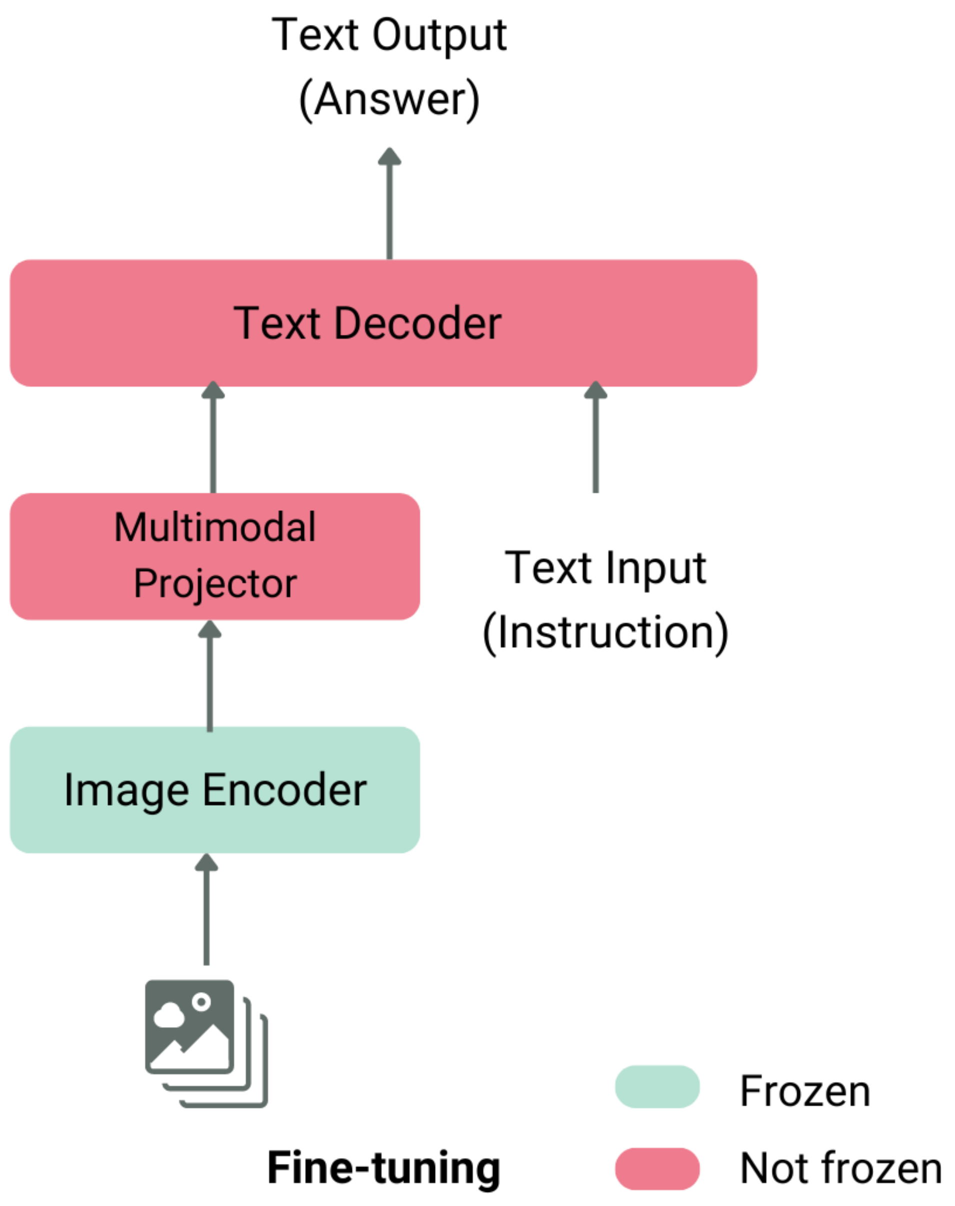
Multilingual Encoder as Plug-in

- use an off-the-shelf multilingual encoder as an plug-in module to map multilingual queries into the English semantic space.

share similar philosophy as VLM







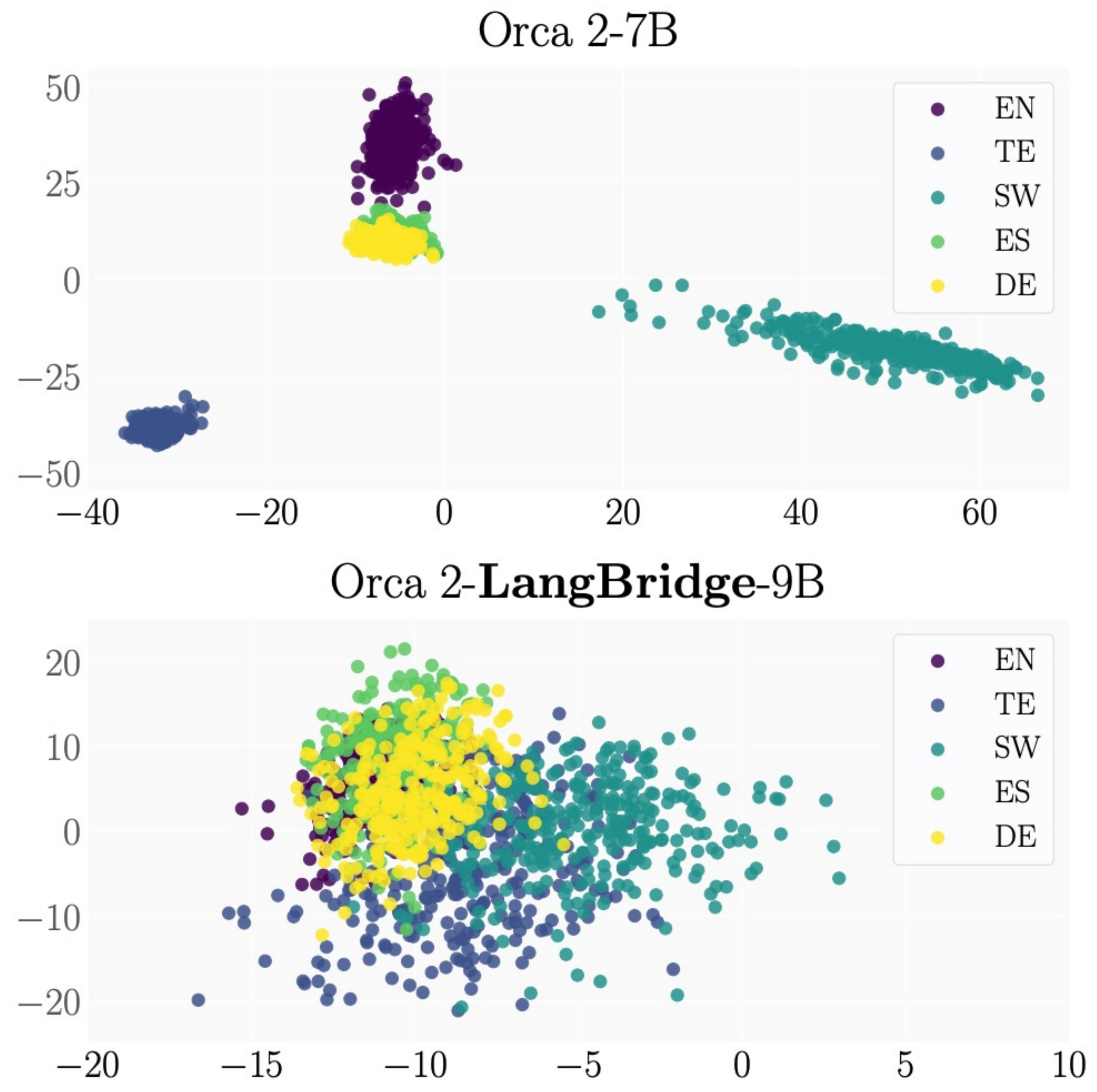
LLaVA (Liu et al.)

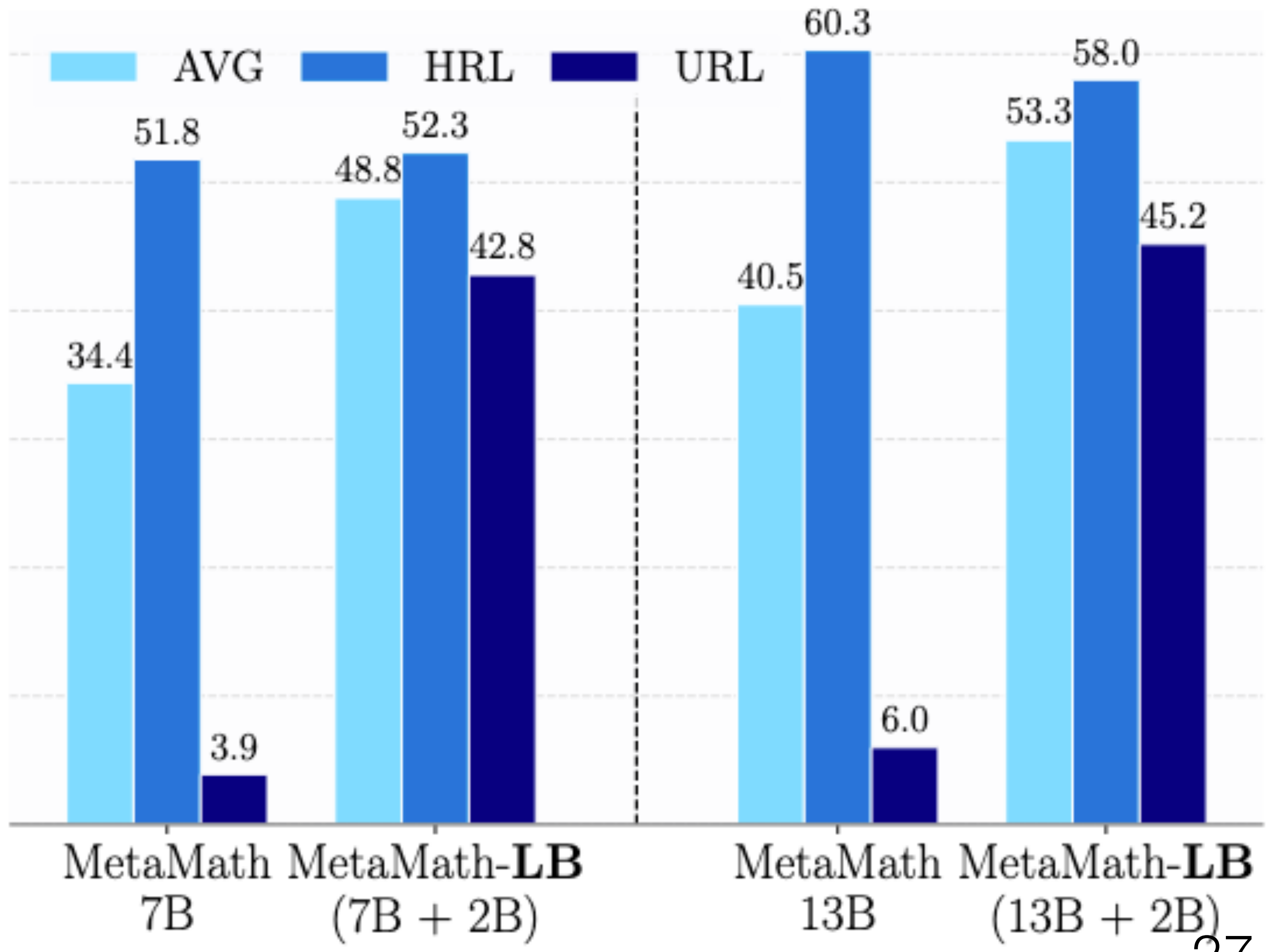
Yoon et al., LangBridge: Multilingual Reasoning Without Multilingual Supervision. ACL'2024.
Huang et al., MindMerger: Efficient Boosting LLM Reasoning in non-English Languages. NeurIPS'2024.
Liu et al., LLaVA: Large Language and Vision Assistant Visual Instruction Tuning. NeurIPS'2023.

Multilingual Encoder as Plug-in

- The multilingual encoder will map multilingual queries into LLM's English representation space.

- The plug-in multilingual encoder significantly narrows the gap between non-English languages and English.
  - does not require any multilingual supervision
  - generalize to multiple languages during test time

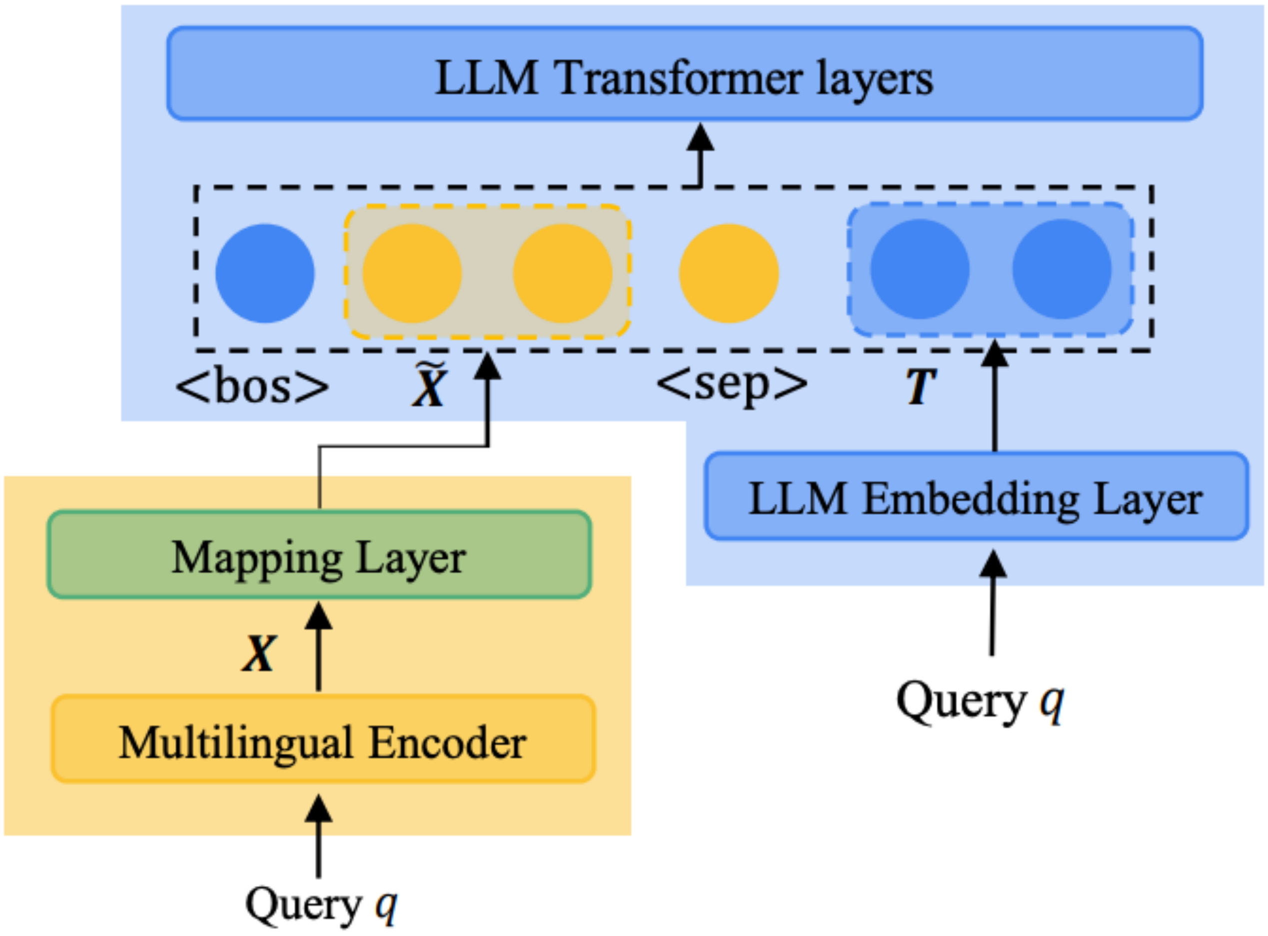




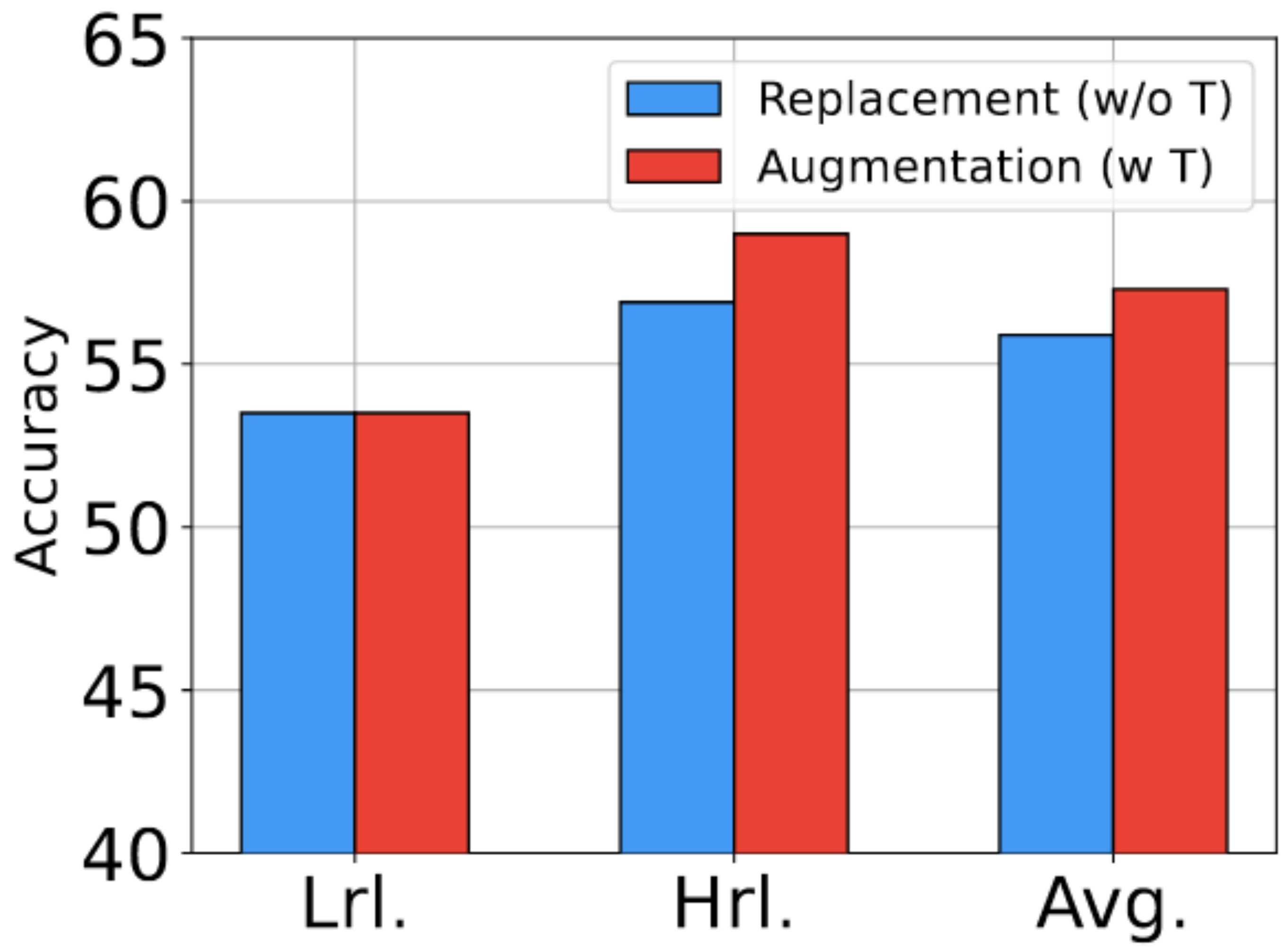
Yoon et al., LangBridge: Multilingual Reasoning Without Multilingual Supervision. ACL'2024.

Multilingual Encoder as Plug-in

- Enhancing input queries with mapped representations is more effective than replacing them.
- Larger encoder often has stronger mapping capability and achieves larger improvements.
- cons: limited in language generation



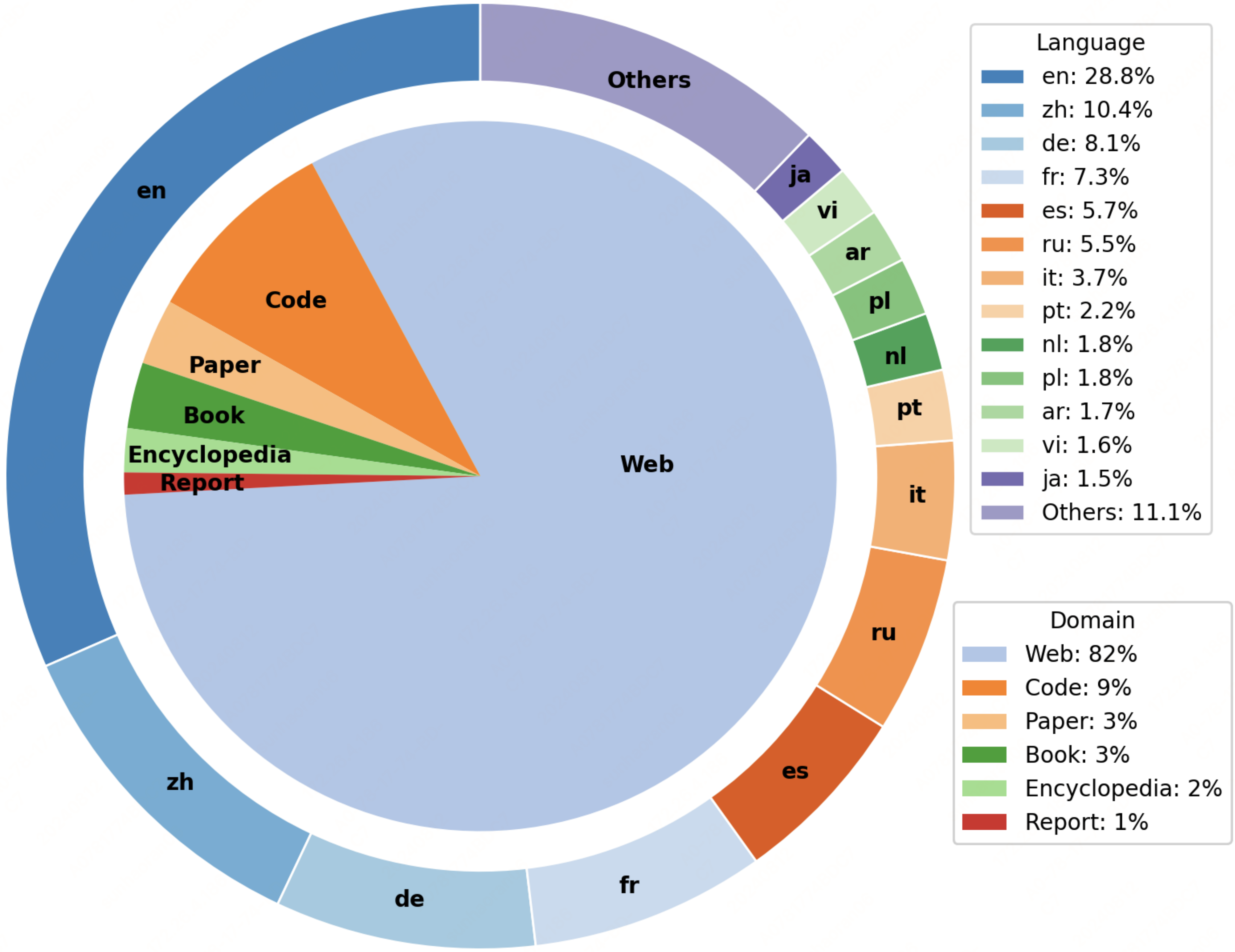
MGSM	# Parm	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Lrl.	Hrl.	Avg.
MindMerger-Soft														
mGPT	1,418 M	19.6	20.4	15.6	42.8	48.0	59.2	59.6	54.0	**61.2**	64.0	18.5	55.5	44.4
mBERT	178 M	30.8	37.6	46.8	50.0	48.8	55.6	52.4	59.6	60.8	66.4	38.4	56.2	50.9
XLM-RoBERTa-large	560 M	44.0	52.4	50.4	52.4	54.0	60.8	58.4	56.8	56.8	66.4	48.9	57.9	55.2
M2M100-418M	282 M	49.2	**52.8**	46.0	48.8	52.4	59.6	58.0	59.2	60.8	65.6	49.3	57.8	55.2
M2M100-1.2B	635 M	49.6	52.4	53.2	52.8	**54.4**	60.0	56.4	60.0	58.0	66.0	51.7	58.2	56.3
NLLB-200-1.3B	766 M	45.6	47.6	**57.6**	**54.4**	52.4	57.2	57.2	**60.8**	60.8	66.8	50.3	58.5	56.2
NLLB-200-3.3B	1,733 M	**52.4**	51.6	53.6	52.8	53.2	60.4	**60.0**	60.4	60.4	**67.6**	52.5	**59.3**	57.2
mT5-large	564 M	40.4	47.2	53.6	47.6	51.6	59.2	55.2	57.6	56.8	66.4	47.1	56.3	53.6
mT5-xl	1,670 M	50.4	**52.8**	57.2	**54.4**	53.6	**61.2**	57.6	**60.8**	58.4	66.8	**53.5**	59.0	**57.3**



Huang et al., MindMerger: Efficient Boosting LLM Reasoning in non-English Languages, NeurIPS'2024.

Balanced Pretraining

- Improving the ratio of other languages by sampling
  - show better alignment among languages
  - fall behind English-centric models

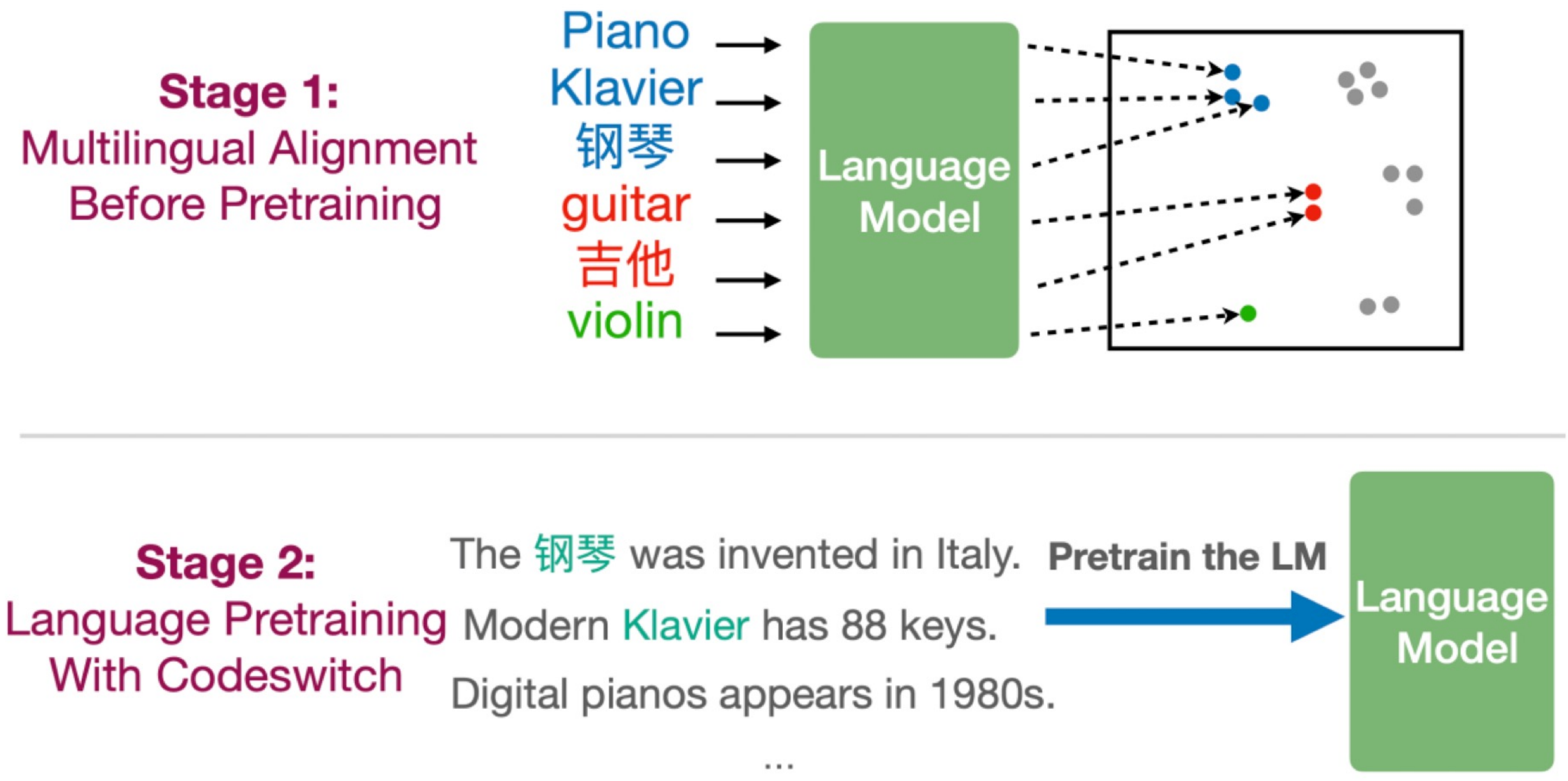


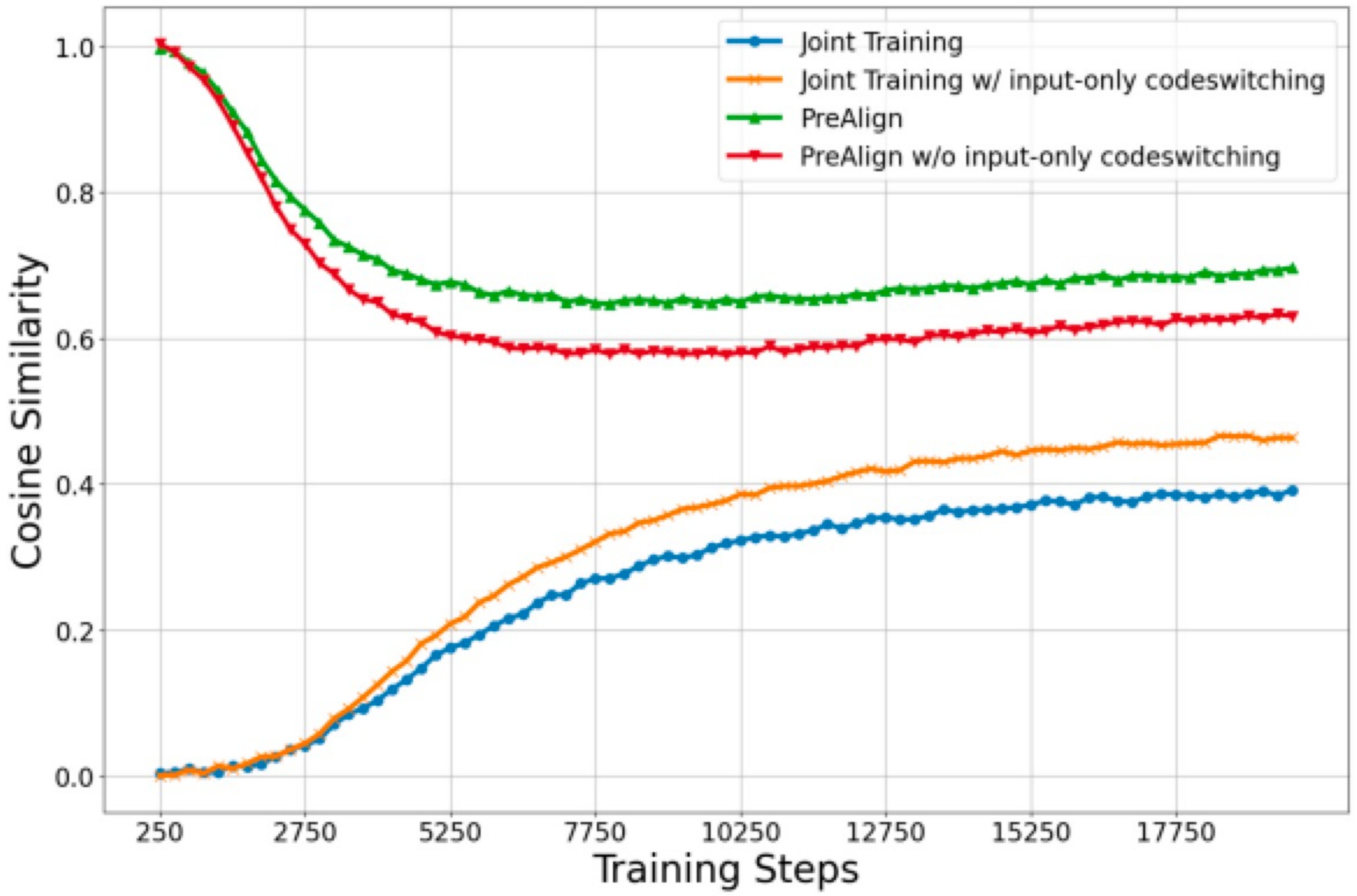
Models	m-ARC (25-shot)	m-Hellaswag (10-shot)	m-MMLU (5-xhot)	XWinograd (5-shot)	XCOPA (0-shot)	XStoryCloze (0-shot)
Llama-2-7B	35.5	48.6	35.4	78.0	58.9	55.6
Mistral-7B-v0.1	**40.7**	**54.5**	**46.7**	**80.5**	55.8	57.2
BLOOM-7B1	31.8	43.4	27.1	70.0	56.9	58.2
PolyLM-13B	30.6	46.0	26.4	73.4	58.9	56.4
LLaMAX2-7B	33.1	50.3	26.7	76.9	54.5	58.8
FuxiTranyu-8B	32.7	51.8	26.6	76.1	**60.5**	**58.9**

Sun et al., FuxiTranyu: A Multilingual Large Language Model Trained with Balanced Data, arxiv'2024.

Early Establishment of Alignment

- PreAlign: pretrain the LLM for language alignment
- The alignment may help multilingualism since early stage of pretraining.

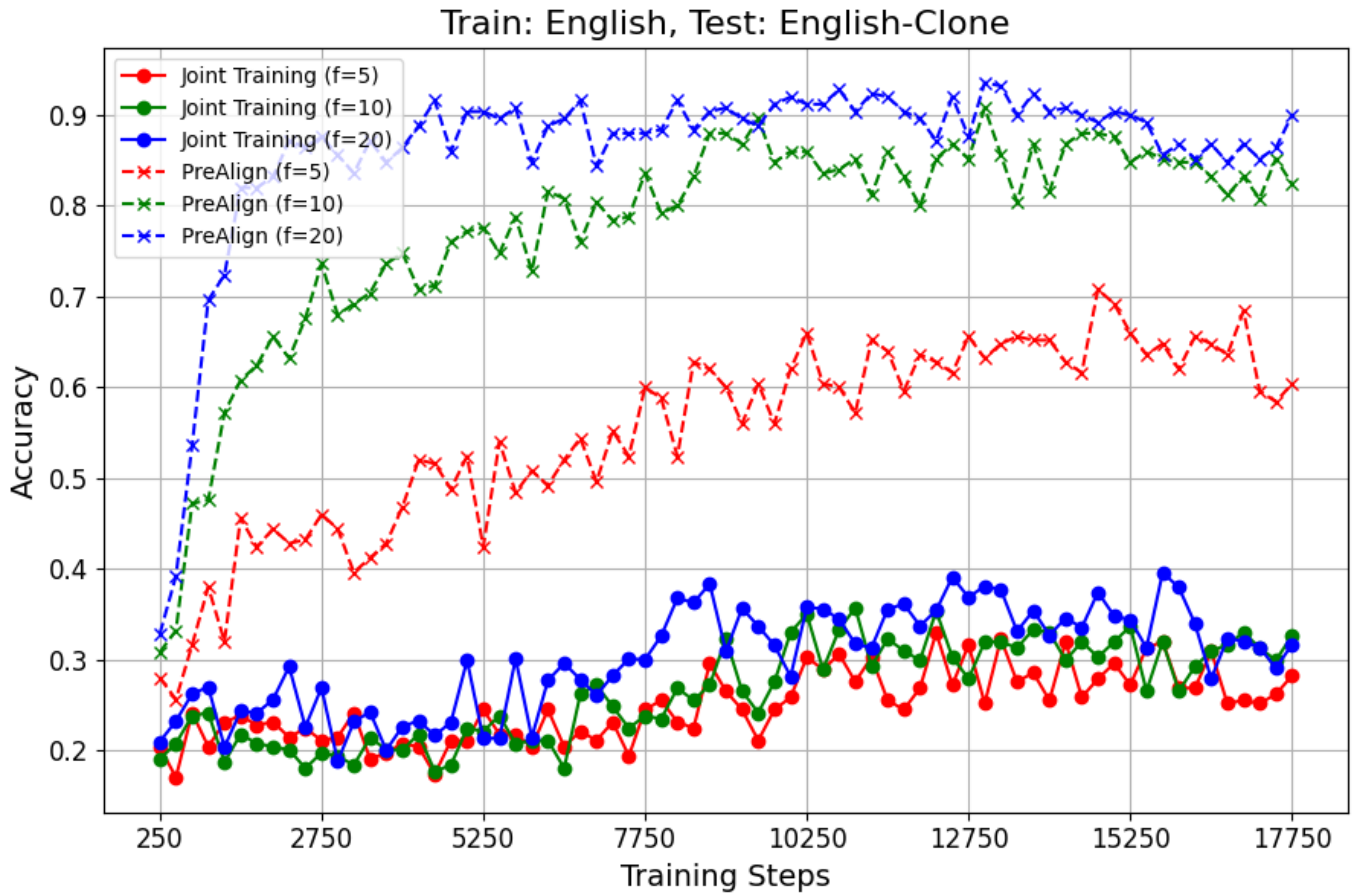




Li et al., PreAlign: Boosting Cross-lingual Transfer by Early Establishment of Multilingual Alignment, EMNLP'2024.

Early Establishment of Alignment

- PreAlign: pretrain the LLM for language alignment
- The alignment may help multilingualism since early stage of pretraining.
  - improves the low-resource language and cross-lingual transfer



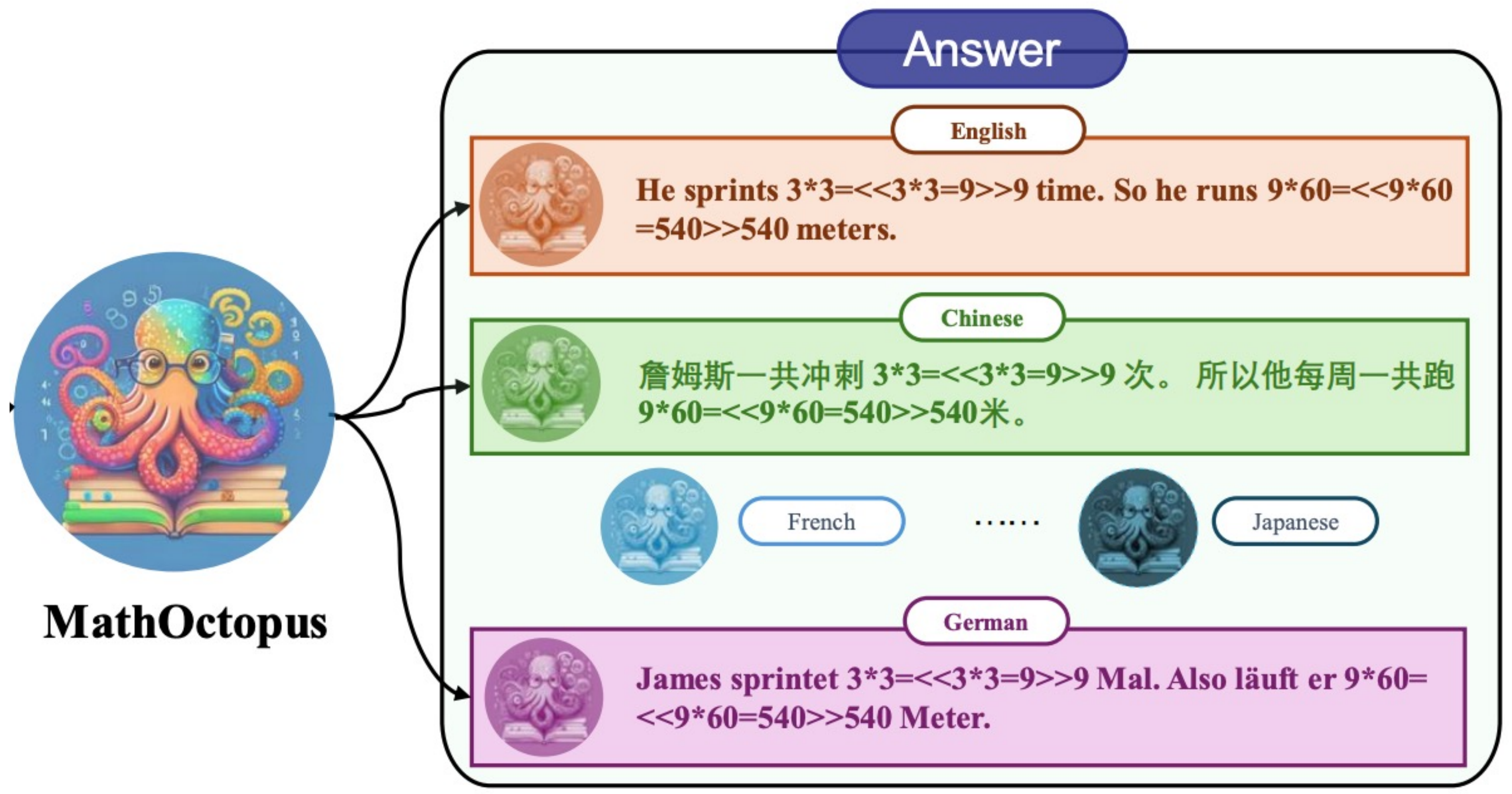
	LM(ppl. ↓)					ZS-CLT(acc. ↑)					CLKA(acc. ↑)			
	En	Zh	De	Ar	Ru	En	Zh	De	Ar	Ru	Zh	De	Ar	Ru
150M														
Joint Training	25.7	99.7	43.5	46.9	49.8	**80.6**	64.6	63.5	58.3	62.0	26.2	25.1	26.8	26.3
PREALIGN	**25.4**	**91.1**	**39.8**	**40.7**	**44.6**	**80.6**	**69.2**	**67.5**	**60.8**	**65.1**	**45.7**	**48.2**	**43.4**	**46.0**
400M														
Joint Training	20.3	79.8	32.5	34.8	39.6	82.3	65.8	65.3	56.9	63.7	37.8	39.5	36.1	37.7
PREALIGN	**19.9**	**75.2**	**28.3**	**30.7**	**33.6**	**82.4**	**70.0**	**69.3**	**65.6**	**68.2**	**63.8**	**66.5**	**64.7**	**63.6**
1.3B														
Joint Training	**15.8**	62.2	24.0	27.7	31.2	**84.3**	70.8	70.6	63.7	68.6	49.6	44.1	45.5	48.0
PREALIGN	16.1	**58.0**	**23.3**	**25.3**	**29.4**	83.9	**74.0**	**72.9**	**68.2**	**71.4**	**71.1**	**73.9**	**72.7**	**72.5**

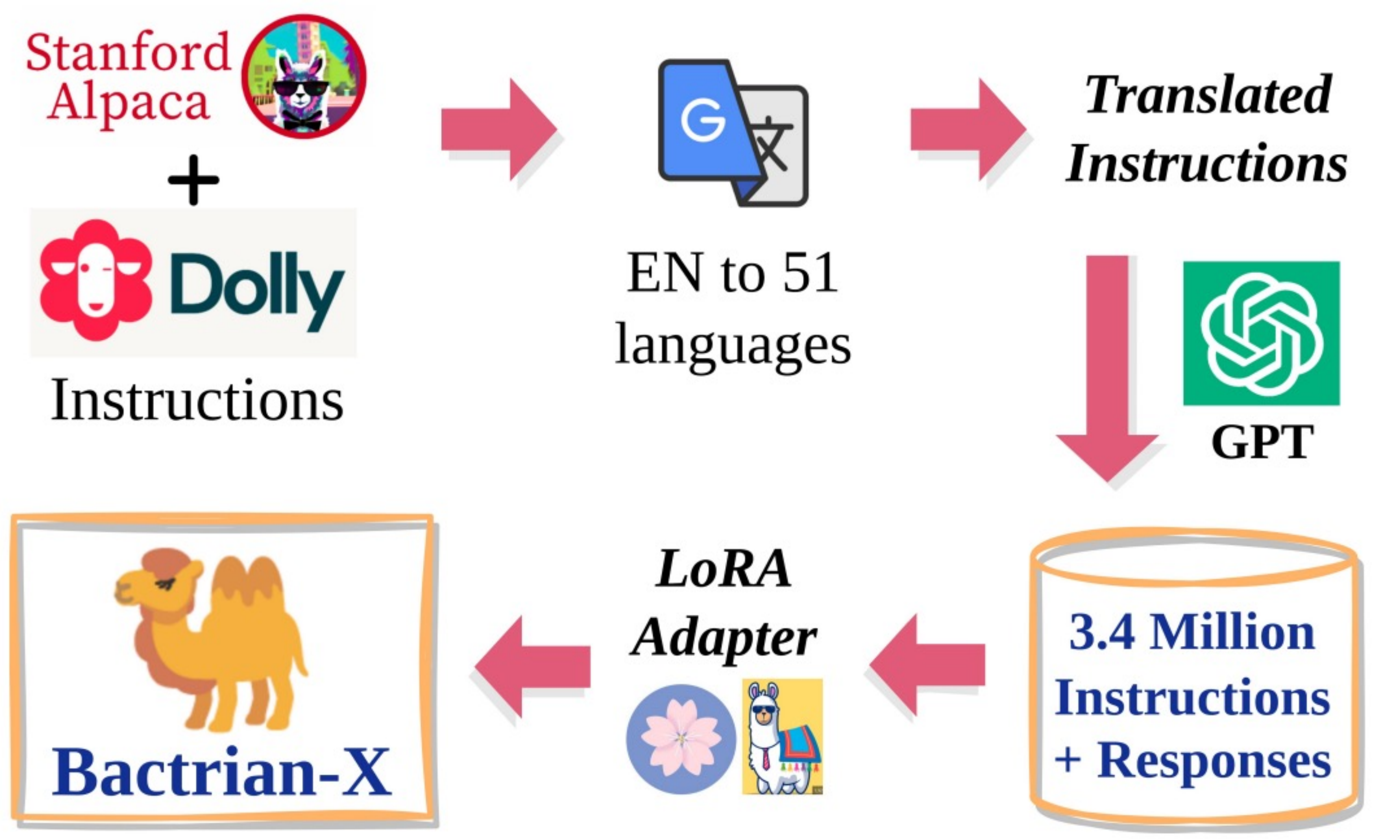
Table 6: Performance of Joint Training and PREALIGN across different scale of models on language modeling, zero-shot cross-lingual transfer (ZS-CLT) and cross-lingual knowledge application (CLKA).

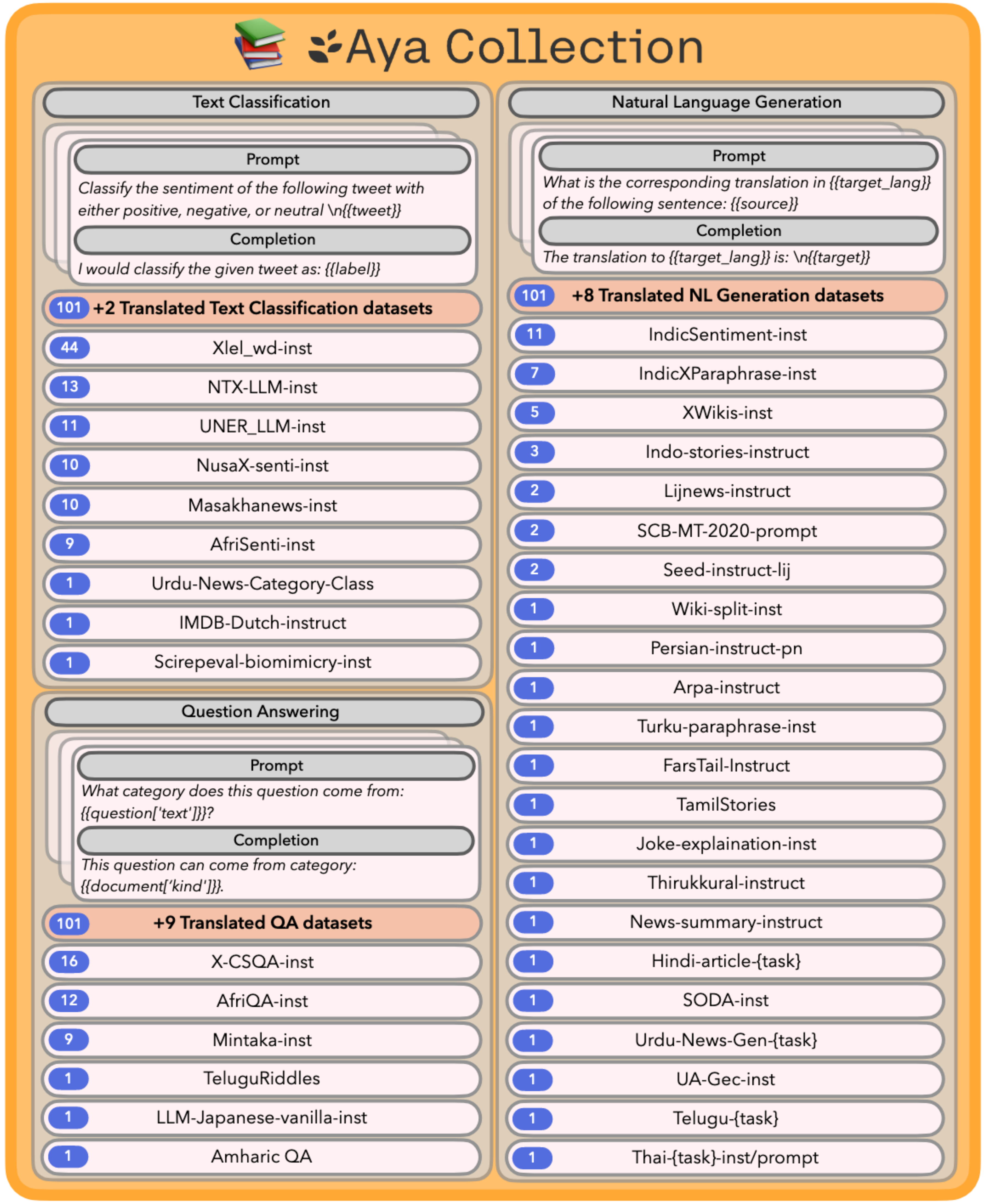
Li et al., PreAlign: Boosting Cross-lingual Transfer by Early Establishment of Multilingual Alignment, EMNLP'2024.

Multilingual Post-training

- One basic idea to enhance non-English performance is to create multilingual post-training data using machine translation.
  - for general task: Aya, Bactrian-X, Okapi
  - for specific task: MathOctopus



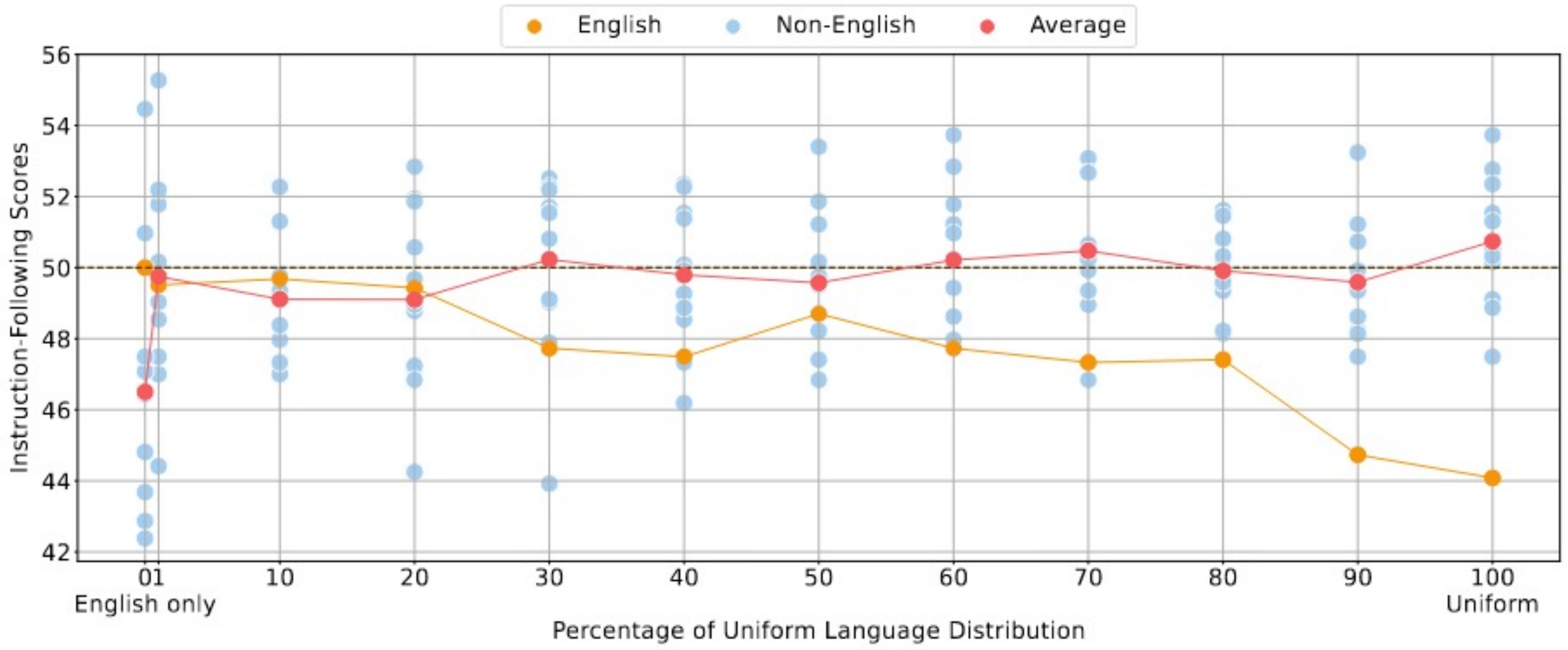


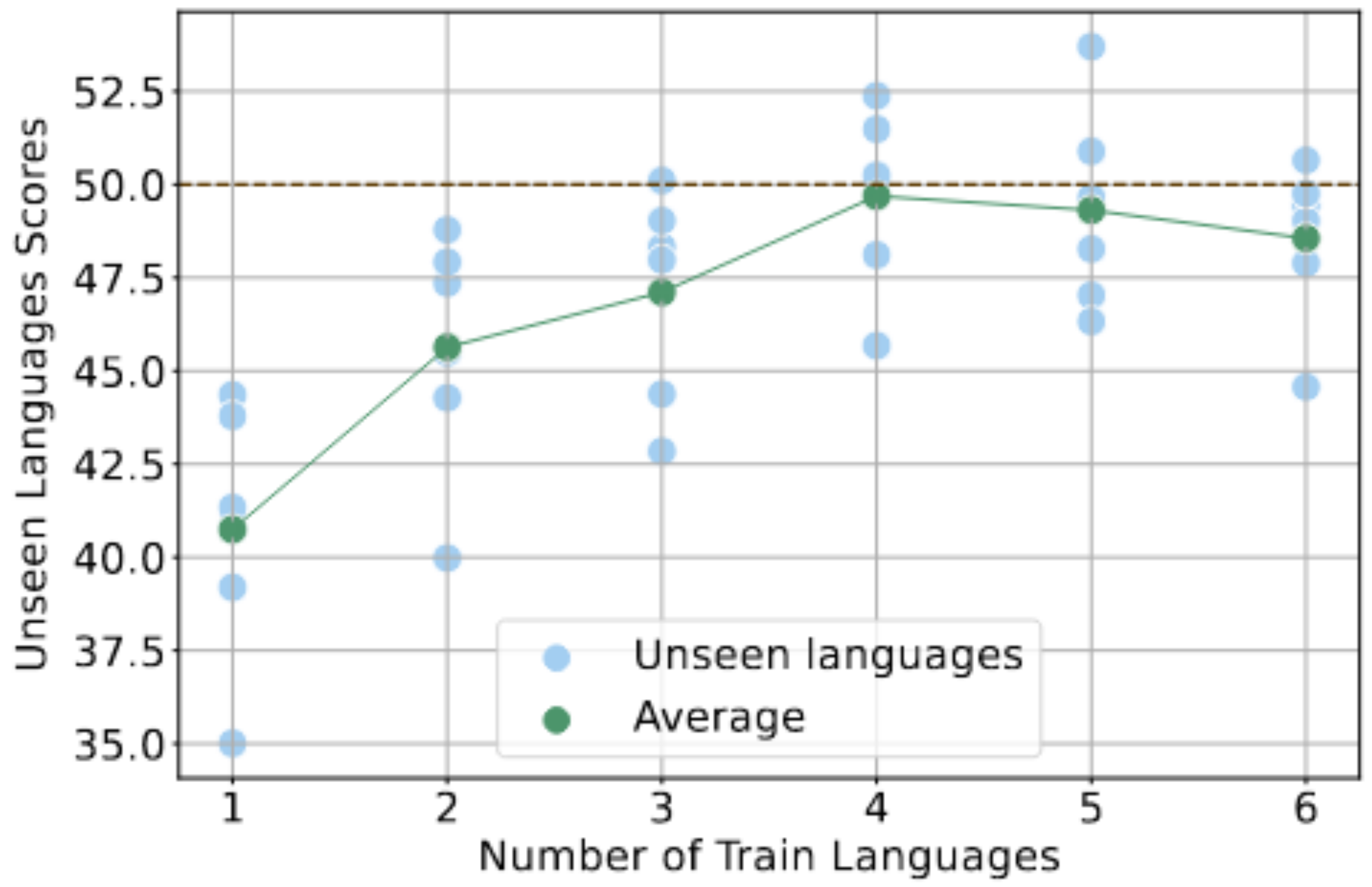


Singh et al., Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. ACL'2024.
Li et al., Bactrian-X : A Multilingual Replicable Instruction-Following Model with Low-Rank Adaptation. arXiv'2023.
Lai et al., Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. EMNLP'2023.
Chen et al., Breaking Language Barriers in Multilingual Mathematical Reasoning: Insights and Observations. arXiv'2023.

Pros of Multilingual Post-training

- performance improvement
  - Improves multilingual performance with limited data (Shaham et al.).
- language generalization
  - Enhances cross-lingual generalization in unseen languages (Shaham et al., Kew et al., Muennighoff et al.).





Muennighoff et al., Crosslingual Generalization through Multitask Finetuning, ACL'2023
Shaham et al., Multilingual Instruction Tuning With Just a Pinch of Multilinguality, ACL'2024
Kew et al, Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed? arXiv'2024

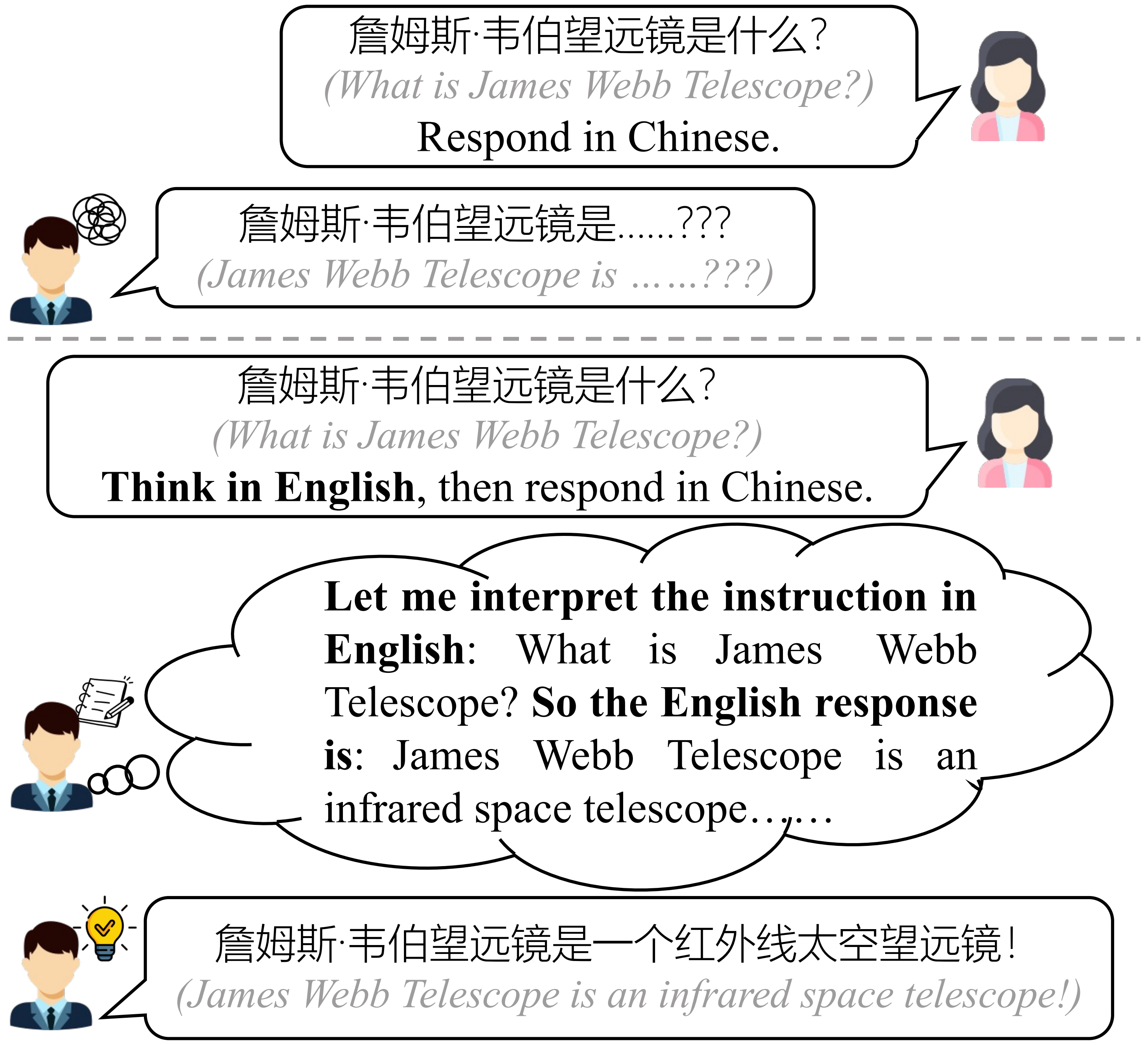
Cons of Multilingual Post-training

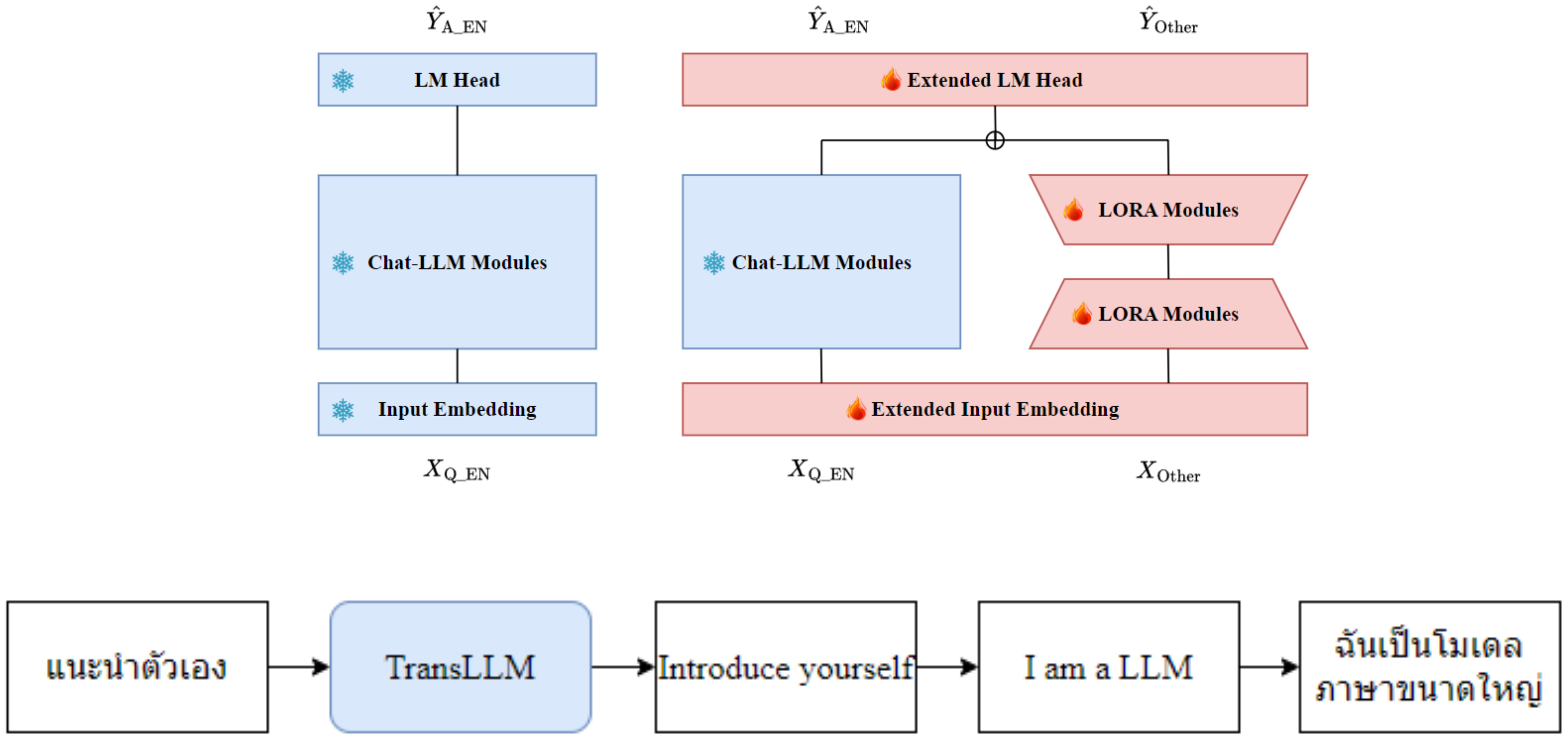
- cross-lingual alignment
  - Multilingual Instruction-tuning can hardly improve cross-lingual consistency and conductivity (Gao et al.).
- data quality
  - Translation engines struggle with lengthy texts with symbols (Zhu et al.).
- annotation cost
  - Translating training data into multiple languages is costly, and evolving datasets quickly make translations outdated.

Gao et al., Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly, NAACL'2024.
Zhu et al., Question Translation Training for Better Multilingual Reasoning, Findings of ACL'2024.

Leveraging Pivot Languages

- It is more challenge to improve advanced abilities, such as instruction following, multi-turn conversation, human alignment, etc.
- Leveraging a pivot language, such as English, improves the process.





Zhang et al., Plug: Leveraging pivot language in cross-lingual instruction tuning, ACL'2024.
Geng et al., Why Not Transform Chat Large Language Models to Non-English? arxiv'2024.

Take-away

- Enabling support for more languages with existing LLMs involves **continue pre-training**: tokenization, data mixture, multilingual curse, etc.
- Multilingualism could also be taken care of since **pretraining**, or even earlier.
- **Post-training** also improves the multilingual ability, but requires more advanced labeled data.

- Further Step:
  - more efficient solution (data, compute)
  - may come from better alignment/pivot

Tutorial Roadmap

- Chapter I: Background
- Chapter II: Observations and Analyses
- Chapter III: Enhancing LLM for More Languages
- **Chapter IV: Aligning Non-English to English**
- Chapter V: Future Challenges

Make the Best Use of LLM's English Expertise

- Explicit Approaches
  - Prompting LLM to think in English (Shi et al., Qin et al.)
  - Prompting LLM to translate the question and answer (Shi et al., Huang et al., Qin et al.)

- Implicit Approaches
  - Eliciting English thinking with translation tasks (Zhu et al.)
  - Improve non-English thinking with English thinking via Preference optimization (She et al.)

- Test-bed
  - mGSM (the multilingual benchmark adopted by most leading LLM teams)

mGSM: Multilingual Mathematical Reasoning

- Task: based on the given math question, predict the numerical answer with multiple reasoning steps.
- Shi et al. extend this to a multilingual task (mGSM).

English

Question: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?

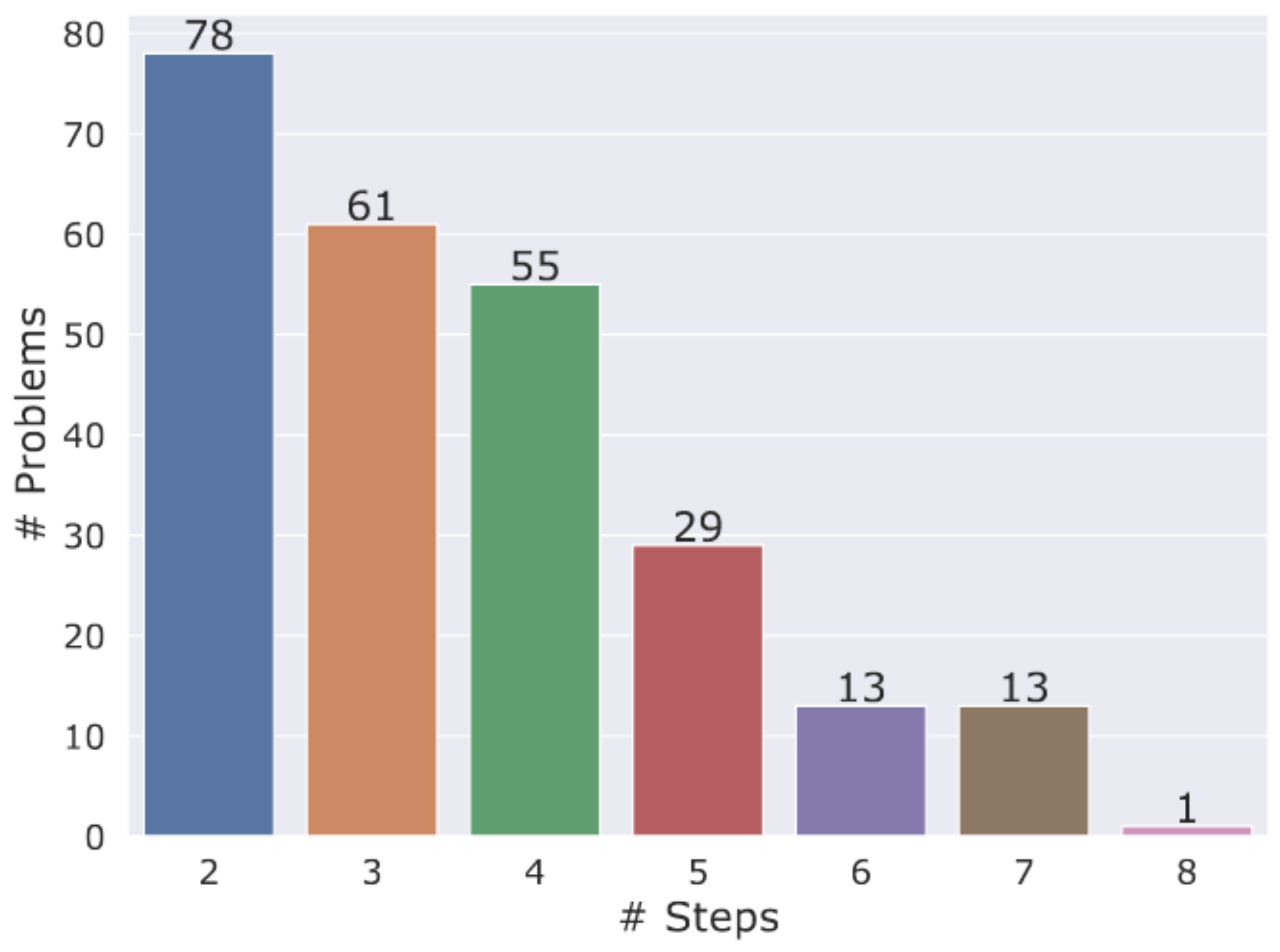
↓ human-translate

German

Frage: Olivia hat 23 US-Dollar. Sie hat fünf Bagels für 3 US- Dollar pro Stück gekauft. Wie viel Geld hat sie übrig?

Chinese

问题： 奥利维亚有 23 美元。她买了五个单价 3 美元的百吉饼。 她还剩多少钱?



Shi et al., Language Models Are Multilingual Chain-Of-Thought Reasoners. ICLR'2023.

Explicit Approach: Ask LLM to Think in English

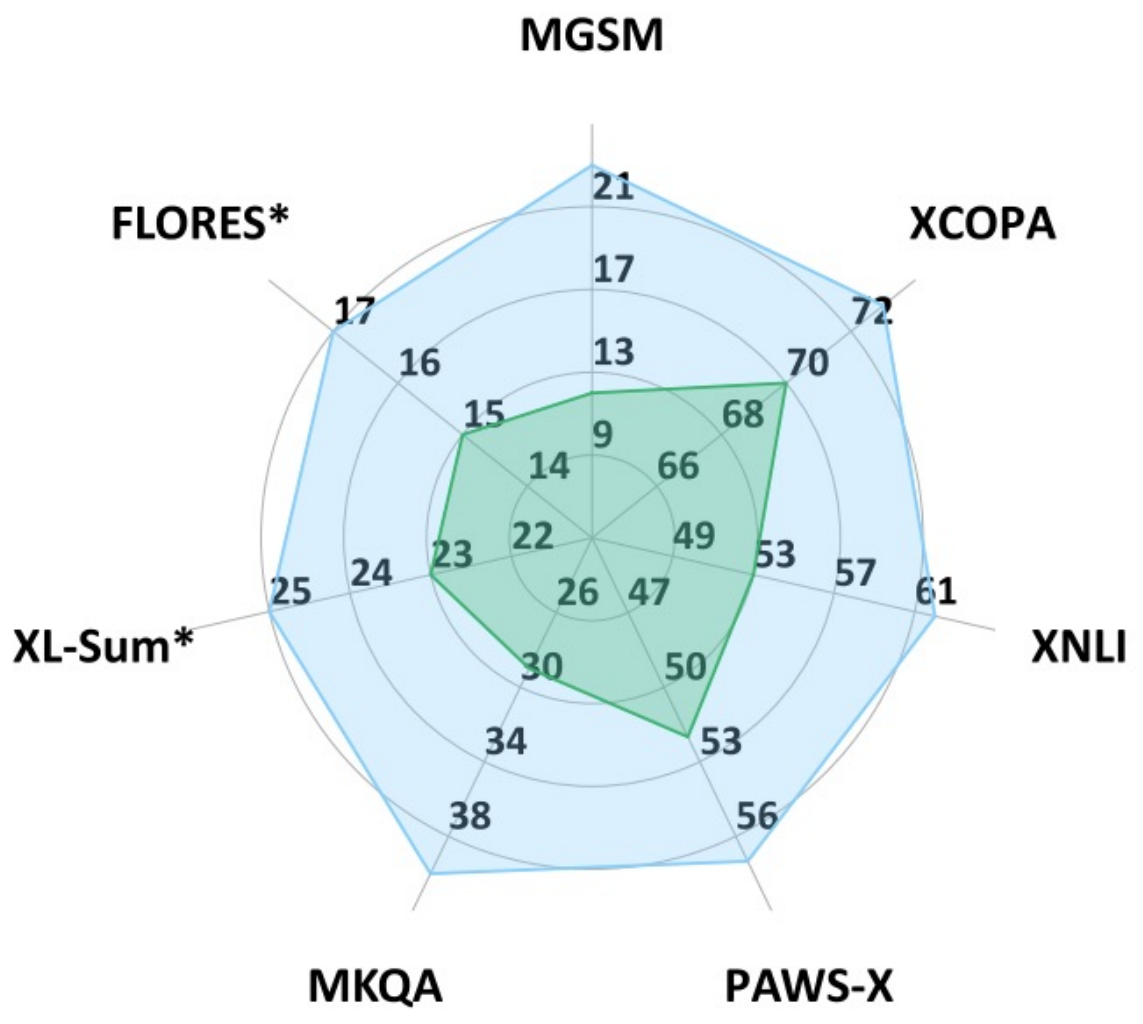
- Intermediate reasoning steps help models achieve substantial reasoning performance gains across all languages.
- Prompting LLM to solve the problem with English CoT
  - English in-context exemplars & English prefix: "Step-by-Step Answer"
  - English CoT consistently lead to competitive or better results than those written in the native language of the question

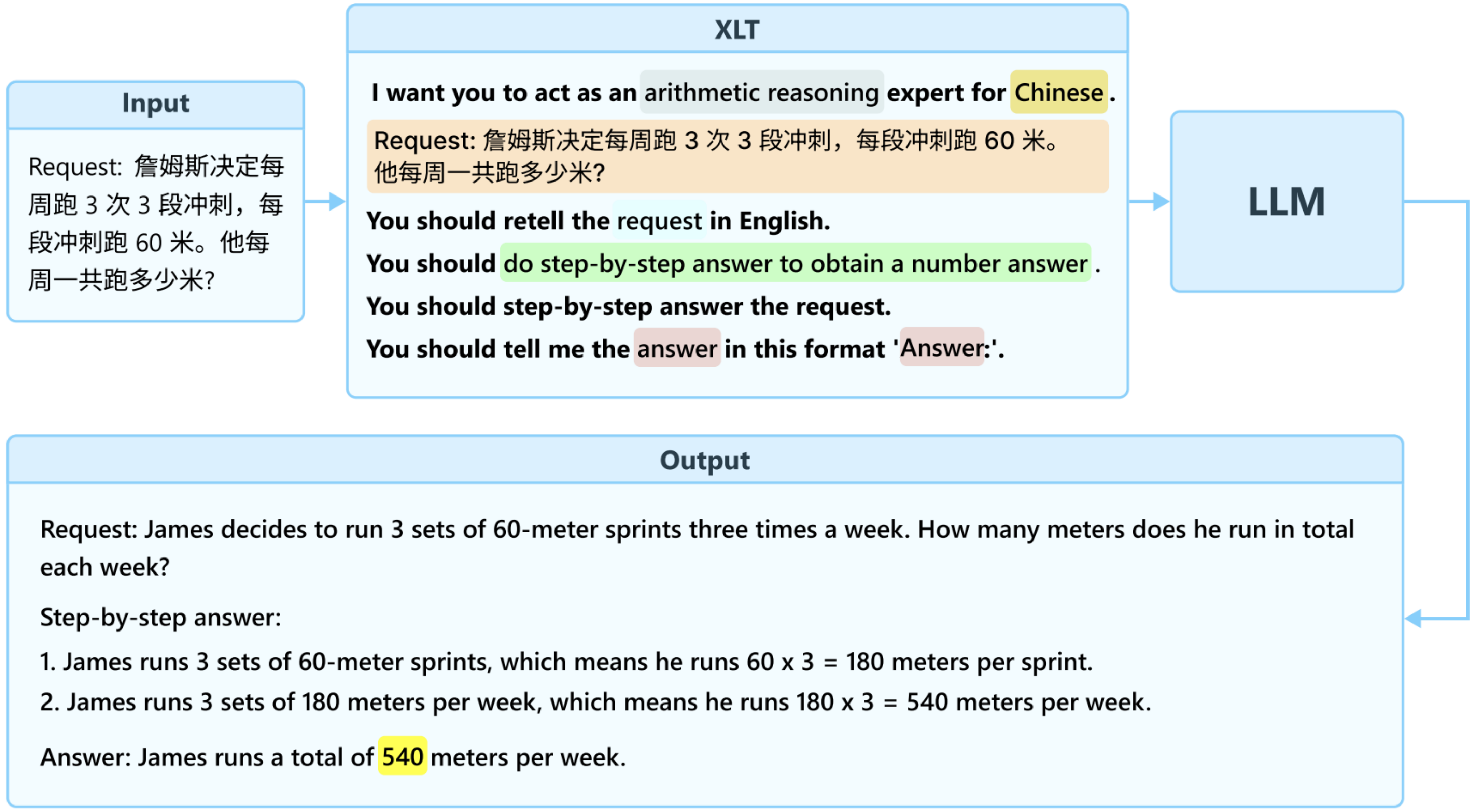
	AVG	HRL	URL	EN	DE	FR	ES	RU	ZH	JA	TH	TE	BN	SW
Lang. Freq. (PaLM, %)	–	–	–	78.0	3.5	3.3	2.1	.53	.40	.38	.04	.02	.006	.005
PaLM-540B														
• DIRECT	18.6	19.3	16.8	22.0	18.8	19.6	20.0	22.0	19.2	16.0	16.8	17.6	17.2	15.6
• NATIVE-COT	48.1	47.9	44.9	**62.4**	49.2	46.4	56.8	48.4	46.8	40.0	52.8	45.6	46.0	35.2
• EN-COT	51.3	52.3	46.8	**62.4**	53.6	51.2	58.0	55.6	46.0	49.6	49.6	46.8	46.4	44.4
• NATIVE-COT-0SHOT	14.4	13.2	7.7	48.0	12.8	12.4	16.8	13.6	10.8	12.8	7.6	6.8	6.8	9.6
• EN-COT-0SHOT	30.8	38.3	15.2	48.0	38.4	36.0	42.4	42.0	35.6	35.2	20.0	10.4	14.0	16.4
• TRANSLATE-EN	**55.0**	**56.3**	**51.2**	**62.4**	**57.2**	**55.2**	**60.0**	**59.6**	**55.6**	**50.0**	**50.8**	**49.6**	**53.2**	**51.2**

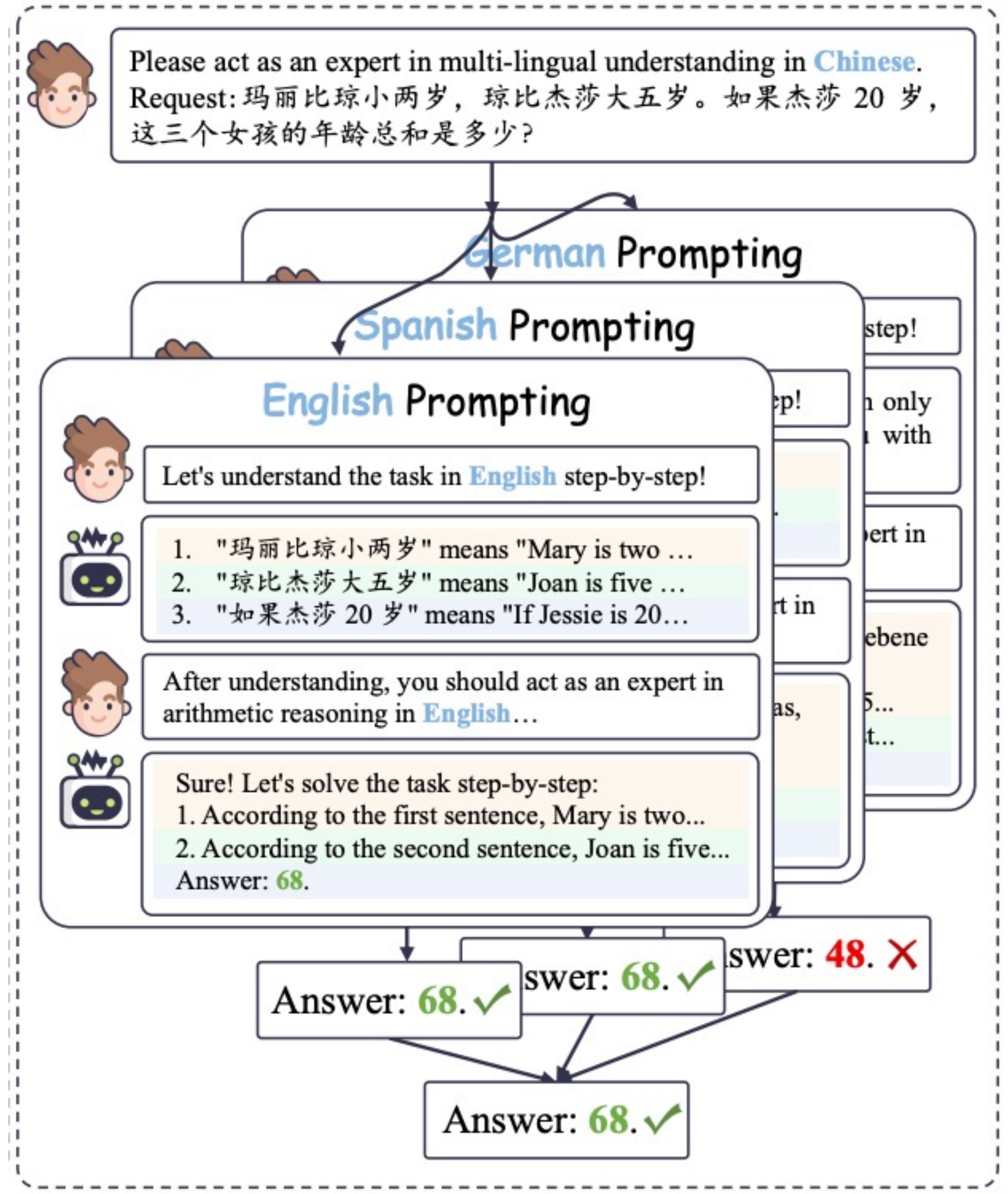
Shi et al., Language Models Are Multilingual Chain-Of-Thought Reasoners. ICLR'2023.

Explicit Approach: Translate-test

- Prompting LLM to translate question into English and answer it with English CoT.
- This increases inference cost and is less effective for LLMs with weak multilingual translation capabilities.



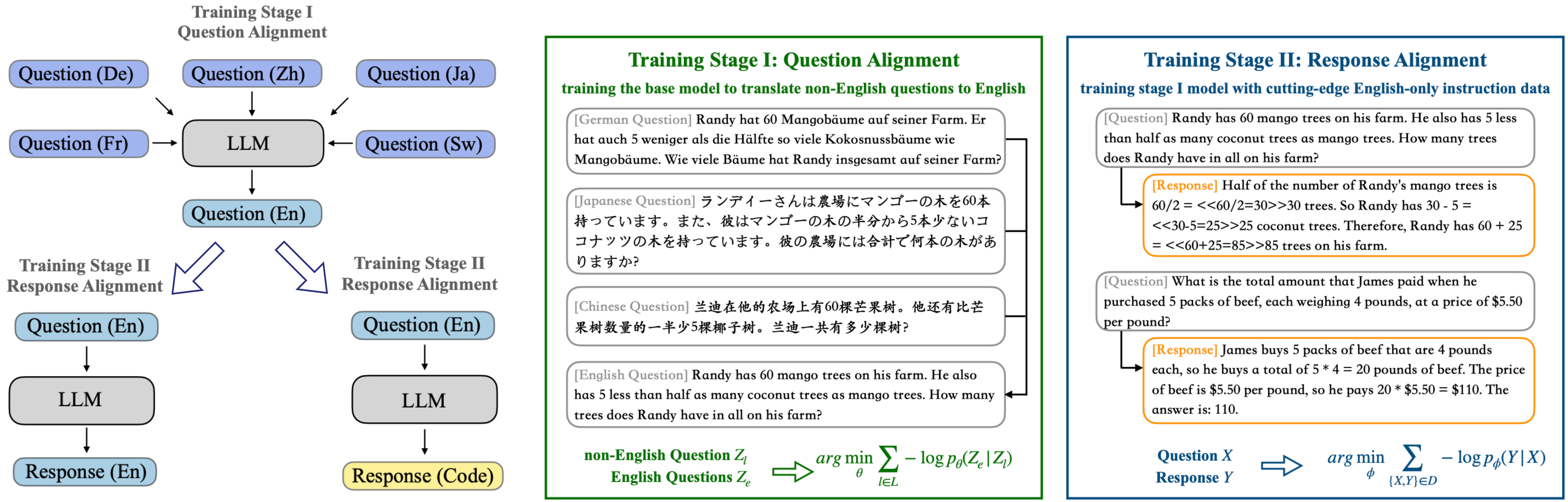




Huang et al., Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting, Findings of EMNLP, 2023
Qin et al., Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages. EMNLP, 2023

Implicit Approach: Question Translation Training

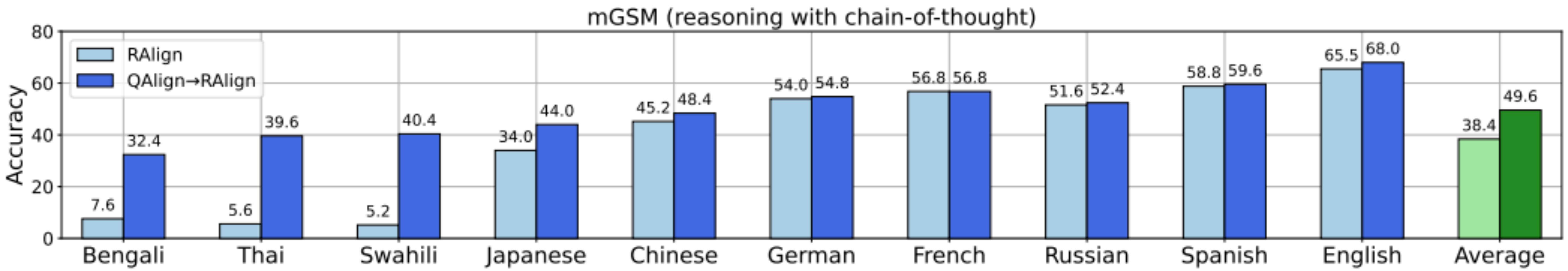
- Training LLM on translating non-English to English strengthens language alignment and implicitly encourages LLM to connect non-English questions to English questions.

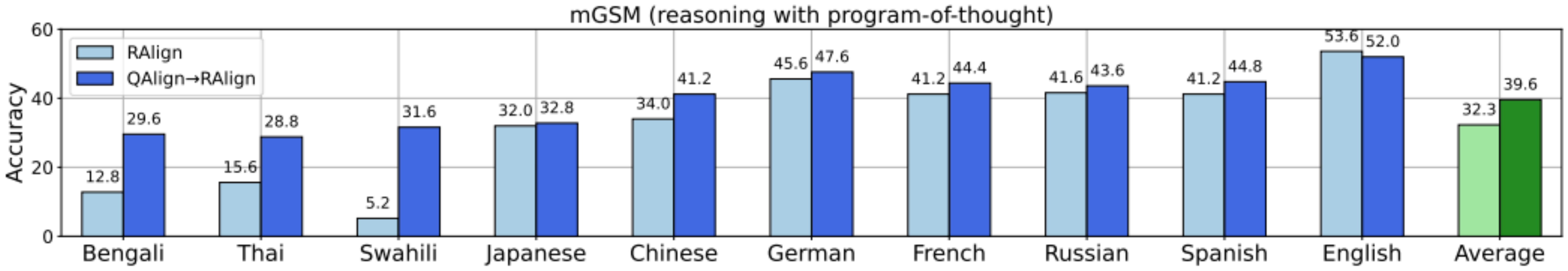


Zhu et al., Question Translation Training for Better Multilingual Reasoning, Findings of ACL'2024.
Zhu et al. The Power of Question Translation Training in Multilingual Reasoning: Broadened Scope and Deepened Insights, arXiv'2024.

Flexible Modular Framework

- The added QAlign stage significantly reduce the gap between non-English languages and English.
- Perform well with both chain-of-thought reasoning and program-of-thought reasoning.

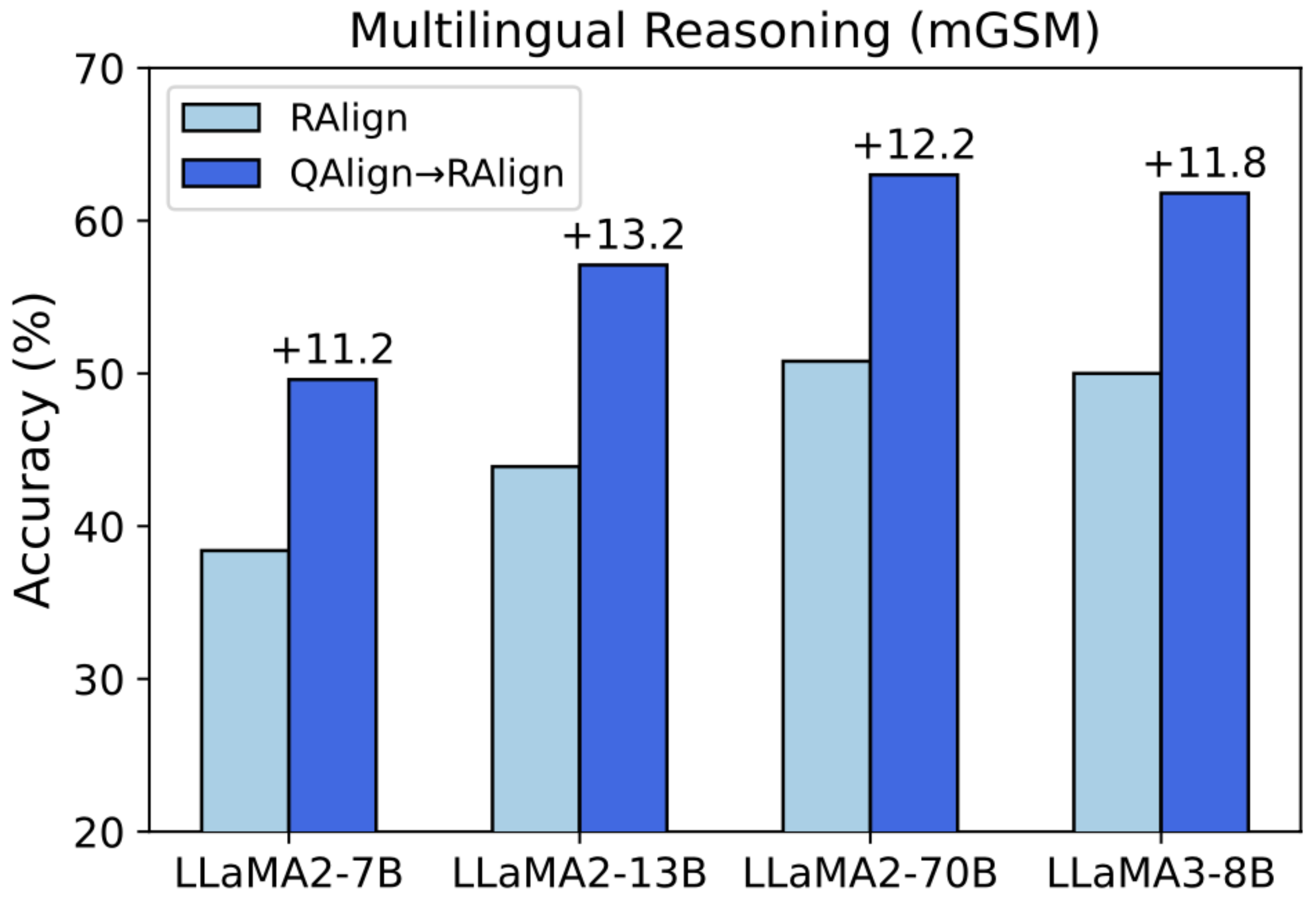




Zhu et al., Question Translation Training for Better Multilingual Reasoning, Findings of ACL'2024.
Zhu et al. The Power of Question Translation Training in Multilingual Reasoning: Broadened Scope and Deepened Insights, arXiv'2024.

Scalable Language Alignment

- The question alignment framework effectively scales to extremely large language models, both dense and sparse.
- Proxy-tuning can quickly extrapolate the results from small models to large models without updating any parameters in the large model.

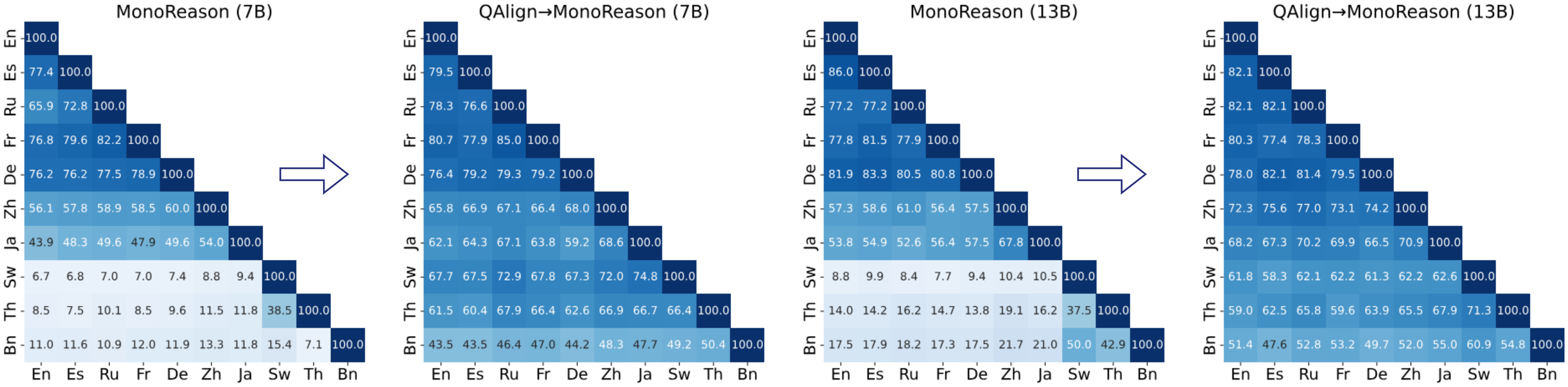


Family	Small tuned $\mathcal{M}^+$	Small untuned $\mathcal{M}^-$	Large untuned $\hat{\mathcal{M}}$	Large tuned $\mathcal{M}$	MGSM Non-En	En	Avg.
LLaMA2	RAlign (13B)	-	-	-	41.2	68.4	43.9
	QAlign→RAlign (13B)	-	-	-	55.7	69.2	57.1
	QAlign→RAlign (13B)	LLaMA2 (13B)	LLaMA2 (70B)	-	60.1	76.8	61.8
LLaMA3	RAlign (8B)	-	-	-	47.3	74.4	50.0
	QAlign→RAlign (8B)	-	-	-	58.4	72.0	59.8
	QAlign→RAlign (8B)	LLaMA3 (8B)	LLaMA3 (70B)	-	**64.0**	**77.2**	**65.4**
Mistral	RAlign (7B)	-	-	-	35.2	70.4	38.7
	QAlign→RAlign (7B)	-	-	-	48.2	70.8	50.4
	QAlign→RAlign (7B)	Mistral (7B)	Mixtral (8×7B)	-	49.4	74.4	51.9
	QAlign→RAlign (7B)	Mistral (7B)	Mixtral (8×22B)	-	**55.6**	**78.0**	**57.9**

Zhu et al., The Power of Question Translation Training in Multilingual Reasoning: Broadened Scope and Deepened Insights, arXiv'2024.
Liu et al., Tuning Language Models by Proxy, COLM'2024.

Consistency across Multilingual Query

- Another evidence of establishing language alignment is the improvement it brings to the consistency of predicted answers against multilingual queries.

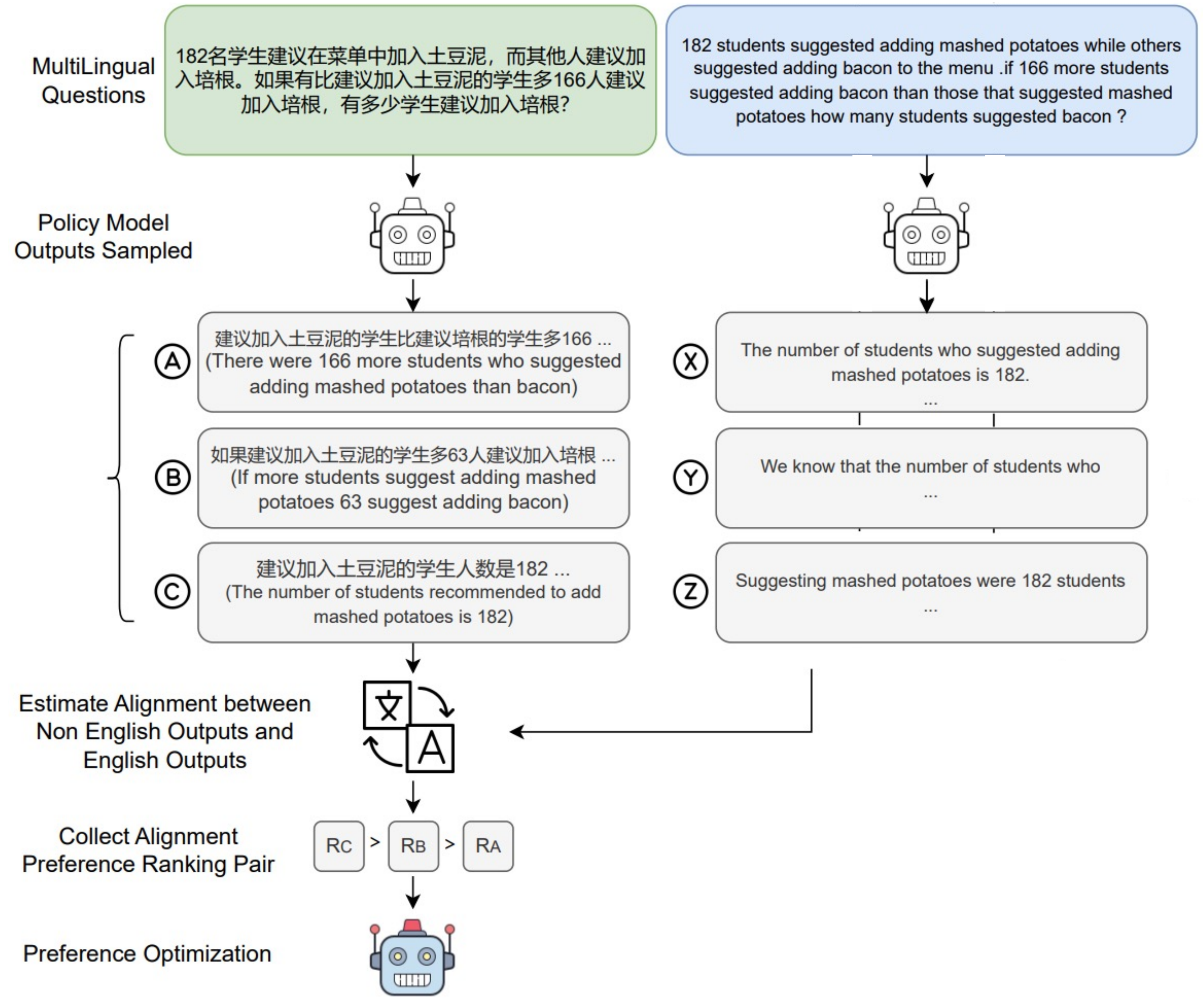


Zhu et al., Question Translation Training for Better Multilingual Reasoning, Findings of ACL'2024.
Zhu et al. The Power of Question Translation Training in Multilingual Reasoning: Broadened Scope and Deepened Insights, arXiv'2024.

Implicit Approach: Preference Optimization

- Improve LLM's multilingual ability with the help of its English thinking, which requires no additional labeling
  - free, internal teacher, which requires no additional labeling
  - step 1: preference estimation
  - step 2: preference optimization

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



She et al., MAPO: Advancing Multilingual Reasoning through Multilingual Alignment-as-Preference Optimization, ACL'2024.

Making Reasoning Path More Consistent

- After the preference optimization, the non-English thinking are more similar to the English thinking.

English Solution	The number of students who suggested mashed potatoes is 182. The number of students who suggested bacon is 182+166=348. ✓
Chinese Solution	**[Before Alignment]** 建议加入土豆泥的学生比建议加入培根的学生多166人，所以两边减去166得到差值为182 - 166 = 16，因此，建议加入培根的学生人数为16。(There are 166 more students recommending mashed potatoes than bacon so subtract 166 from both sides to get a difference of 182-166 = 16. Therefore, the number of students recommended to join Bacon is 16.) ✗ **[PPL]:** 2.65 **[After Alignment]** 建议加入土豆泥的学生人数是182。建议加入培根的学生人数是182 + 166 = 348。(The number of students who suggested mashed potatoes is 182. The number of students who suggested bacon is 182+166=348.) ✓ **[PPL]:** 0.97

Model	**Bn**	**Th**	**Sw**	**Ja**	**Zh**	**Ru**	**De**	**Es**	**Fr**	**En**	**Avg**
MathOctopus 7B	29.2	33.6	36.4	35.2	39.2	38.8	44.8	42.4	43.2	52.0	39.5
+ m-RFT	25.6	31.2	28.8	34.0	39.2	36.0	34.8	34.4	36.4	43.2	34.4
+ MAPO-DPO(ours)	30.8	38.0	37.6	45.2	47.2	42.0	45.2	43.2	40.8	45.6	41.6
MetaMathOctopus 7B	25.6	42.8	36.4	40.0	46.4	46.8	49.6	54.4	46.4	66.4	45.5
+ m-RFT	23.2	33.6	34.0	34.0	47.2	43.2	45.6	47.6	44.8	62.8	41.6
+ MAPO-DPO(ours)	36.0	44.8	44.8	47.6	55.2	53.6	53.6	56.8	52.4	70.8	51.6

She et al., MAPO: Advancing Multilingual Reasoning through Multilingual Alignment-as-Preference Optimization, ACL'2024.

Take-away

- Comparing to multilingual post-training, which requires extensive data labeling, leveraging English abilities seems to be a more efficient solution.
  - close-source models only allow explicit approaches
  - implicit approaches pushes open-source models to a new height.

- Further Step:
  - Using English v.s. Using Native Language
  - More general solution that generalize across tasks.
  - Implicit solution that does not affect user experience.

Tutorial Roadmap

- Chapter I: Background
- Chapter II: Observations and Analyses
- Chapter III: Enhancing LLM for More Languages
- Chapter IV: Aligning Non-English to English
- Chapter V: Future Challenges

What's Next?

- Developing multilingual system for real-world applications

  - evaluation📏: from specific tasks to general tasks

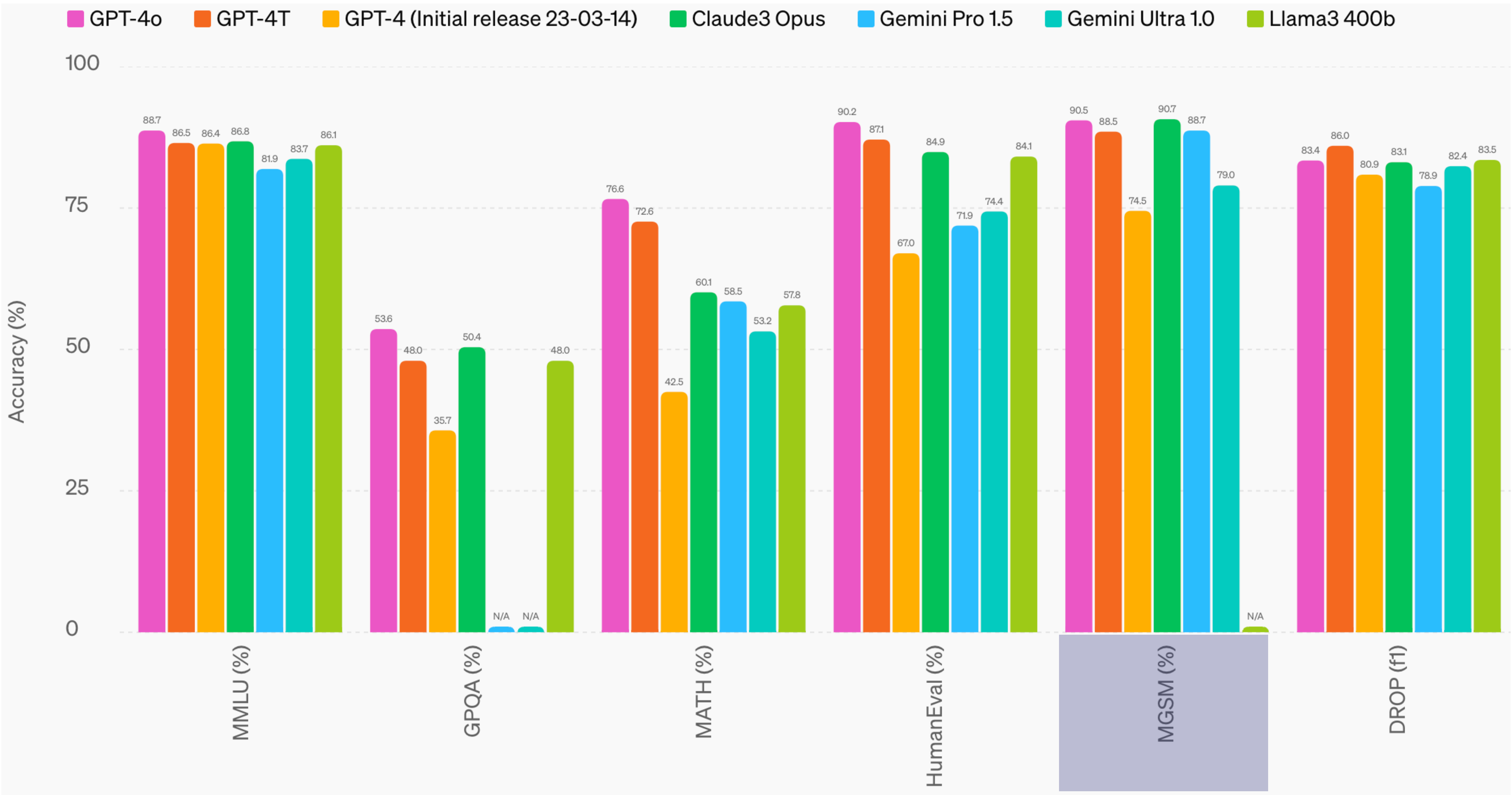
  - data📊: from basic data mixing to strategic data mixing

  - model🧭: from action model to reward model

  - culture⛵: from fully sharing to selective sharing

From Specific Task to General Tasks

- Math reasoning is still far away from real-world applications.
- Developing more powerful multilingual systems requires the curation of reliable and comprehensive benchmark.



GPT-4o

Category	Benchmark
General	MMLU (5-shot) MMLU (0-shot, CoT) MMLU-Pro (5-shot, CoT) IFEval
Code	HumanEval (0-shot) MBPP EvalPlus (0-shot)
Math	GSM8K (8-shot, CoT) MATH (0-shot, CoT)
Reasoning	ARC Challenge (0-shot) GPQA (0-shot, CoT)
Tool use	BFCL Nexus
Long context	ZeroSCROLLS/QuALITY InfiniteBench/En.MC NIH/Multi-needle
Multilingual	MGSM (0-shot, CoT)

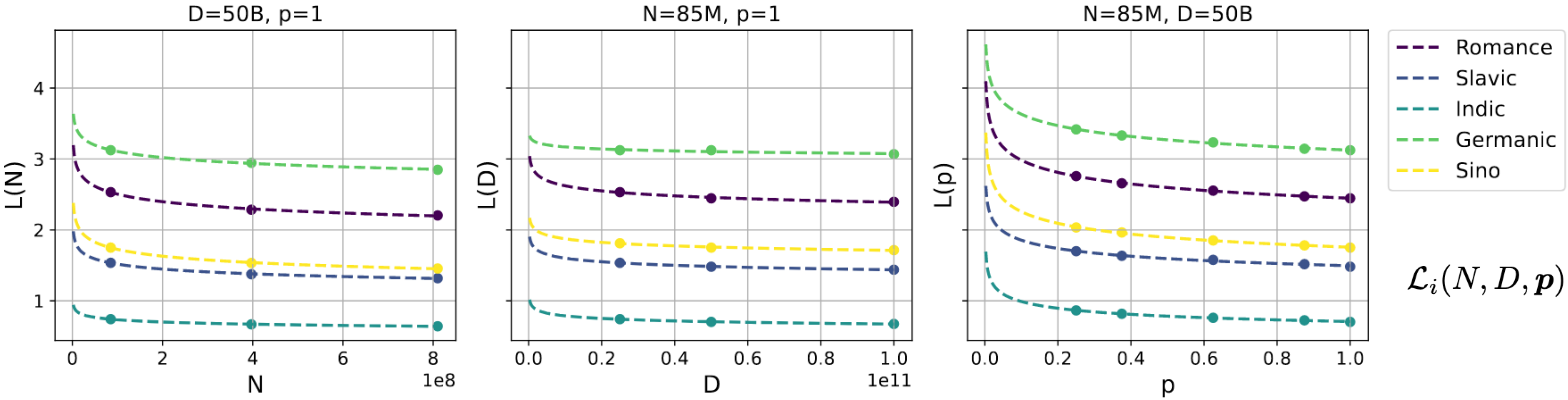
Llama3

Capability	Benchmark
General Reasoning	**BigBench - Hard:** A subset of harder tasks from Big Bench. (Srivastava et al., 2022; Suzgun et al., 2022)
	DROP: Reading comprehension & arithmetic. (Metric: F1-Score). (Dua et al., 2019)
	MMLU: Multiple-choice questions in 57 subjects (professional & academic). (Hendrycks et al., 2021a)
	Hellaswag (Zellers et al., 2019)
Coding	**HumanEval** chat preamble* (Metric: pass rate). (Chen et al., 2021)
	Natural2Code chat preamble* (Metric: pass rate).
Multilinguality	**WMT23:** sentence-level machine translation (Metric: BLEURT). (Tom et al., 2023)
	MGSM: multilingual math reasoning. (Shi et al., 2023a)

Gemini-1.5

From Basic Mixing to Strategic Mixing

- Estimating the optimal data mixture recipe is one of the key problems in multilingual research.
- He et al. pioneered the formulation of a multilingual scaling law.
  - The cross-entropy loss (L) is related to model size (N), dataset size (D), and sampling ratios for different language families (p).

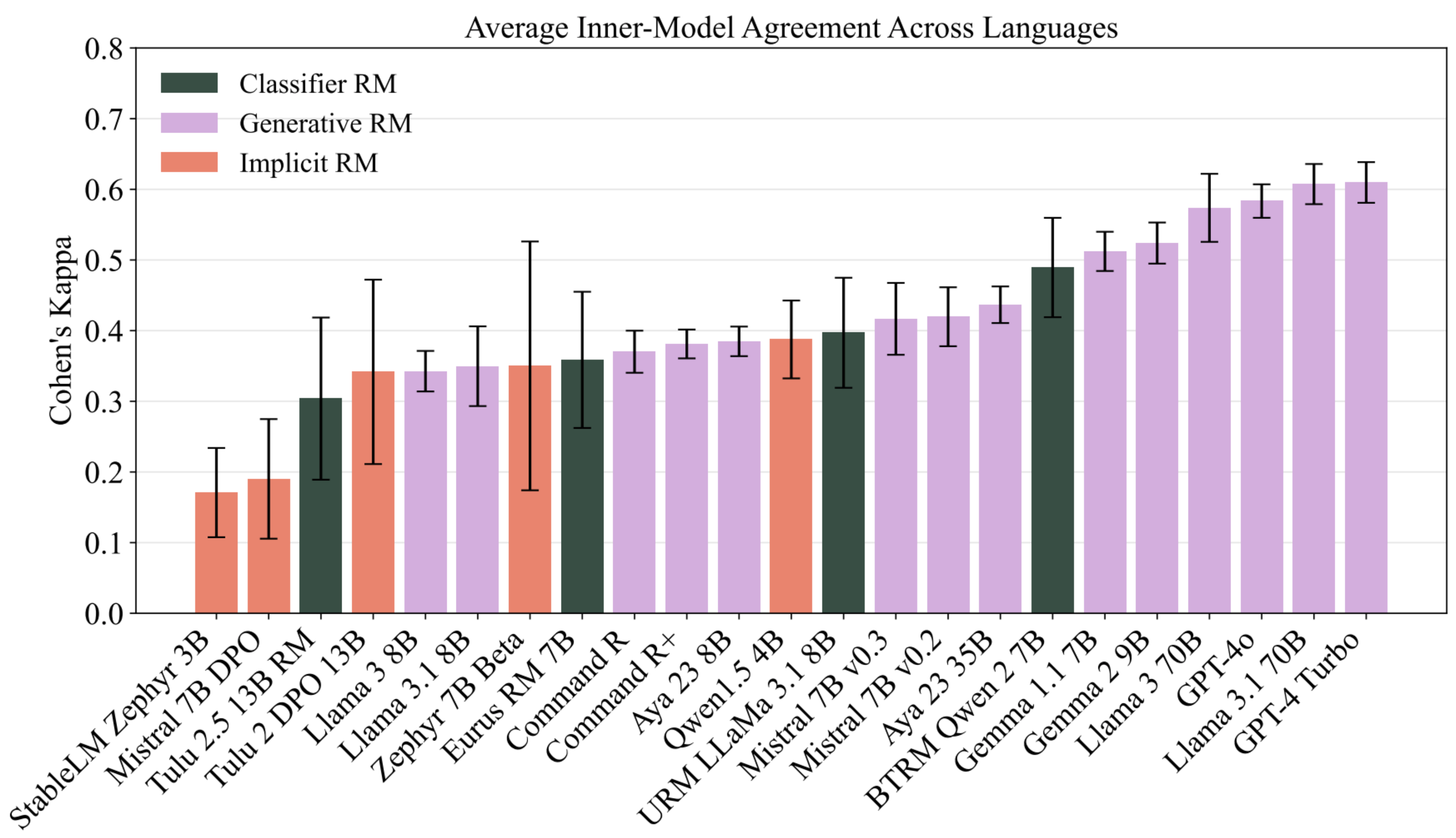


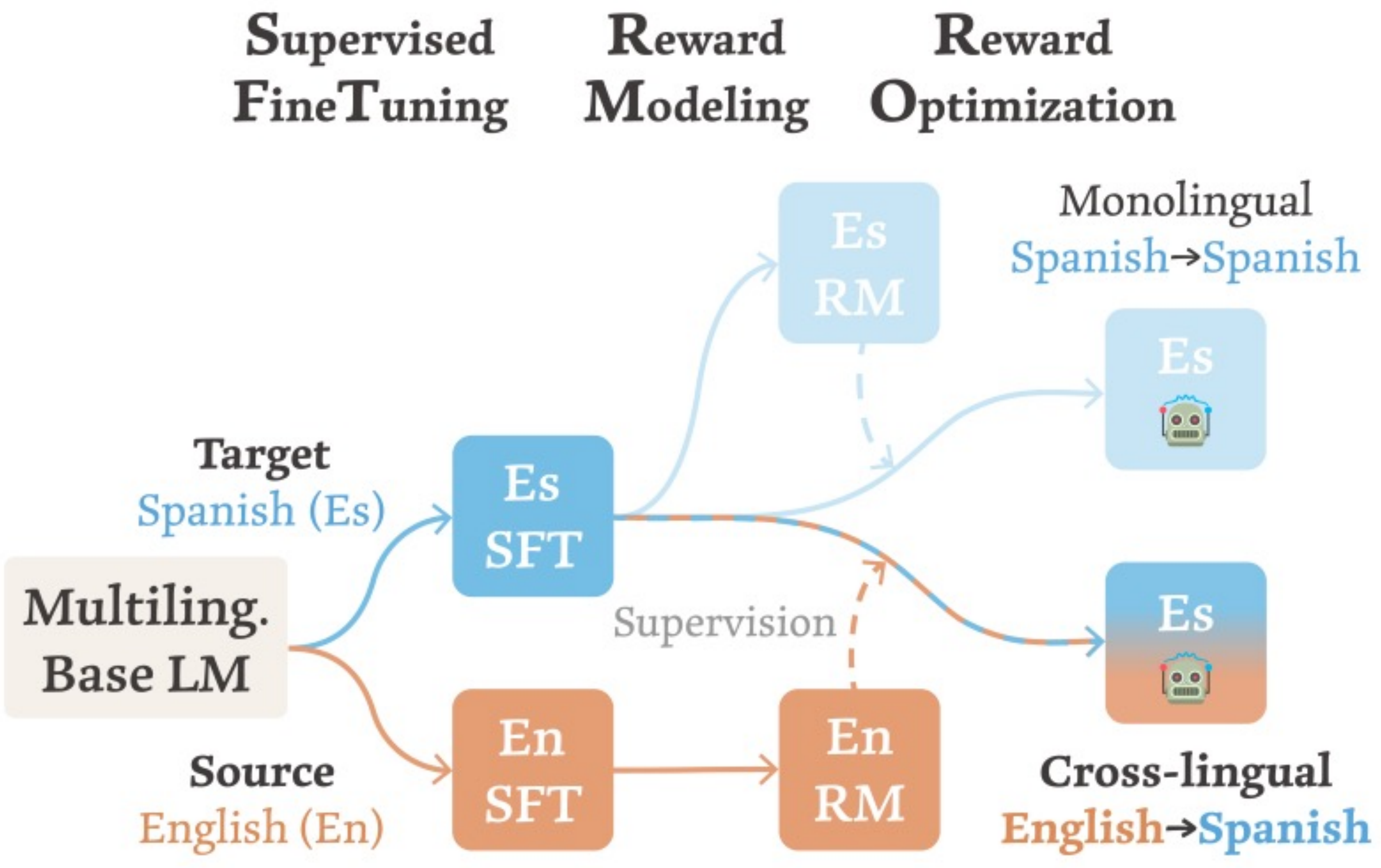
$$\mathcal{L}_i(N, D, \boldsymbol{p}) = \left(E_i + \frac{A_i}{N^{\alpha_i}} + \frac{B_i}{D^{\beta_i}} \right) p_i^{-\gamma_i}$$

He et al., Scaling Laws for Multilingual Language Models, arXiv'2024.

From Action Model to Reward Model

- Reward model is becoming increasingly important
  - LLM-as-a-judge
  - self-improvement
  - test-time scaling
- Multilingual Reward Model
  - Again, reward model face challenges in multilingual context.
  - Is it possible to adapt English reward model to multilingual scenarios?





Wu et al. Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment. arXiv'2024.
Gureja et al. M-REWARDBENCH: Evaluating Reward Models in Multilingual Settings. arXiv'2024.

From Fully Sharing to Selective Sharing

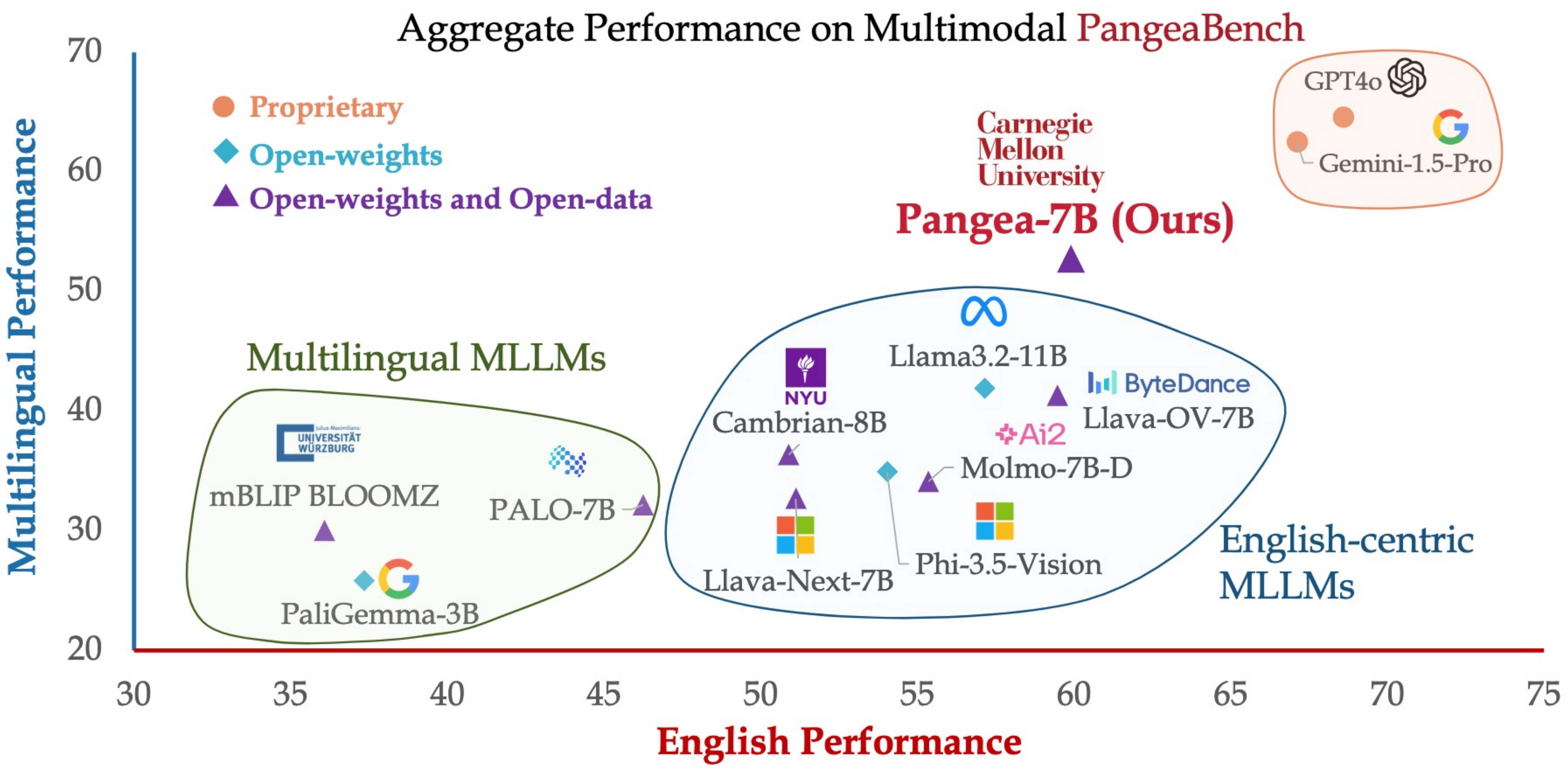
- Not all capabilities/knowledge should be shared across languages, as transferring English proficiency may introduce English bias.
  - For example, dragons symbolize different meanings in different cultures.

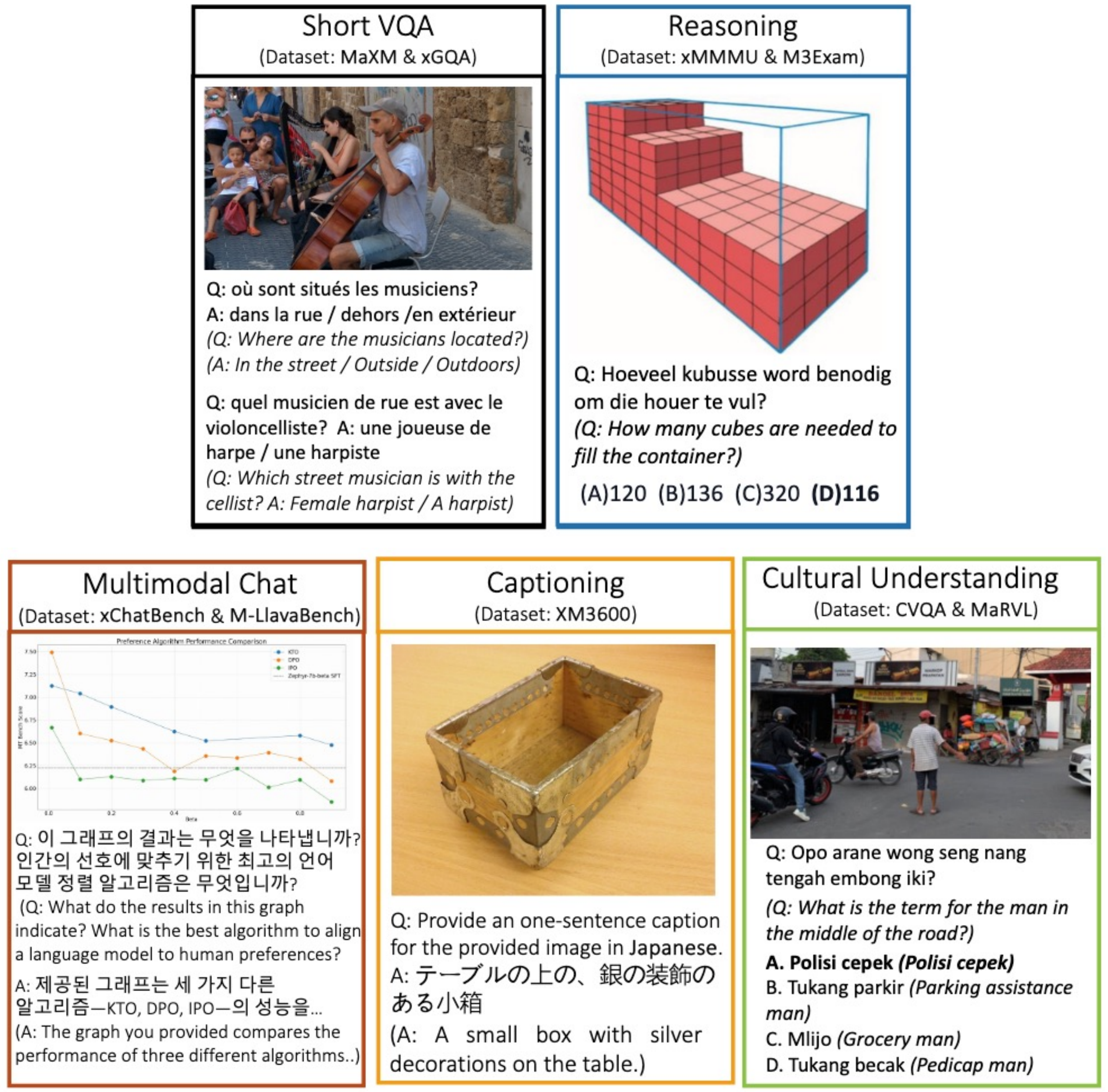
Chinese dragon

western dragon

From Single-Modality to Multi-Modality

- Adding new modalities, such as vision, will definitely makes LLM more capable.
- How will different modalities interact, and how will the added modality impact the model's multilingual capabilities?





Yue et al. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages. arXiv'2024.

Conclusion

- Although LLMs have become highly capable, their multilingual performance remains uneven.
- Breaking the language barriers may be essential for fair-usage of LLMs.
- Progress have been made in understanding and improving the multilingual process of LLMs.
- But still more challenges ahead!
  - Knowledge, Reasoning, Alignment.

Thanks for your attention!

huangsj@nju.edu.cn
zhuwh@smail.nju.edu.cn