Towards Scaling Large Language Models to 1000 Languages

Challenges and Advances

Lei Li

Carnegie Mellon University

November 6, 2024


MARCO POLO
Travels between 1271 and 1295
Venice
Constantinople (Istanbul)
Mediterranean Sea
Black Sea
Bulgar
SIBERIA
Sarai
Aral Sea
Caspian Sea
Ayas
Erzurum
Acre
Jerusalem
Tabriz
Samarkand
Bukhara
HINDU KUSH MTS.
PAMIRS
Kashgar
TAKLA MAKAN DESERT
Khotan (Hotan)
IRAN
Persian Gulf
ARABIA
AFGHANISTAN
Hormuz
HIMALAYAS
INDIA
ETHIOPIA
ARABIAN SEA
GOBI DESERT
Ningxia
Suzhou
Lanzhou
GANSU
CHINA
Shangdu
Dadu (Beijing)
JAPAN
YELLOW SEA
Quinsay (Hangzhou)
Quanzhou
PACIFIC OCEAN
Yunnan
Tagaung
SOUTH CHINA SEA
CHAMPA (VIETNAM)
Nicobar Islands
Malay Peninsula
Ceylon (Sri Lanka)
Sumatra
INDIAN OCEAN
0 500 1000 mi
0 750 1500 km
© Encyclopædia Britannica, Inc.

Breaking Language Barriers

Cultural Communication

Education

Medical care

Tourism

Business&trade

AI Translation has increased international trade by 10%

http://pubsonline.informs.org/journal/mnsc

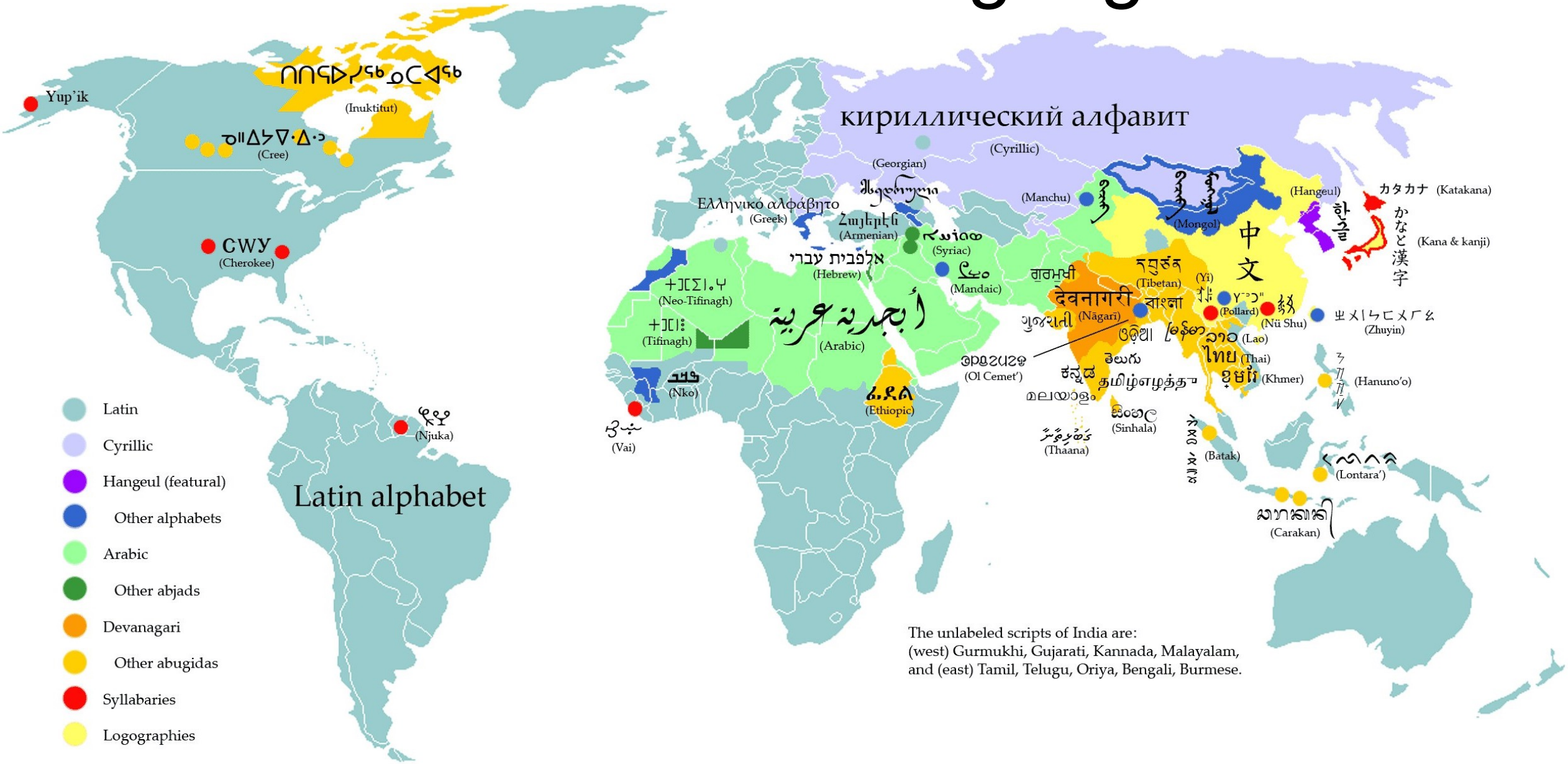
MANAGEMENT SCIENCE
Vol. 65, No. 12, December 2019, pp. 5449–5460
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform

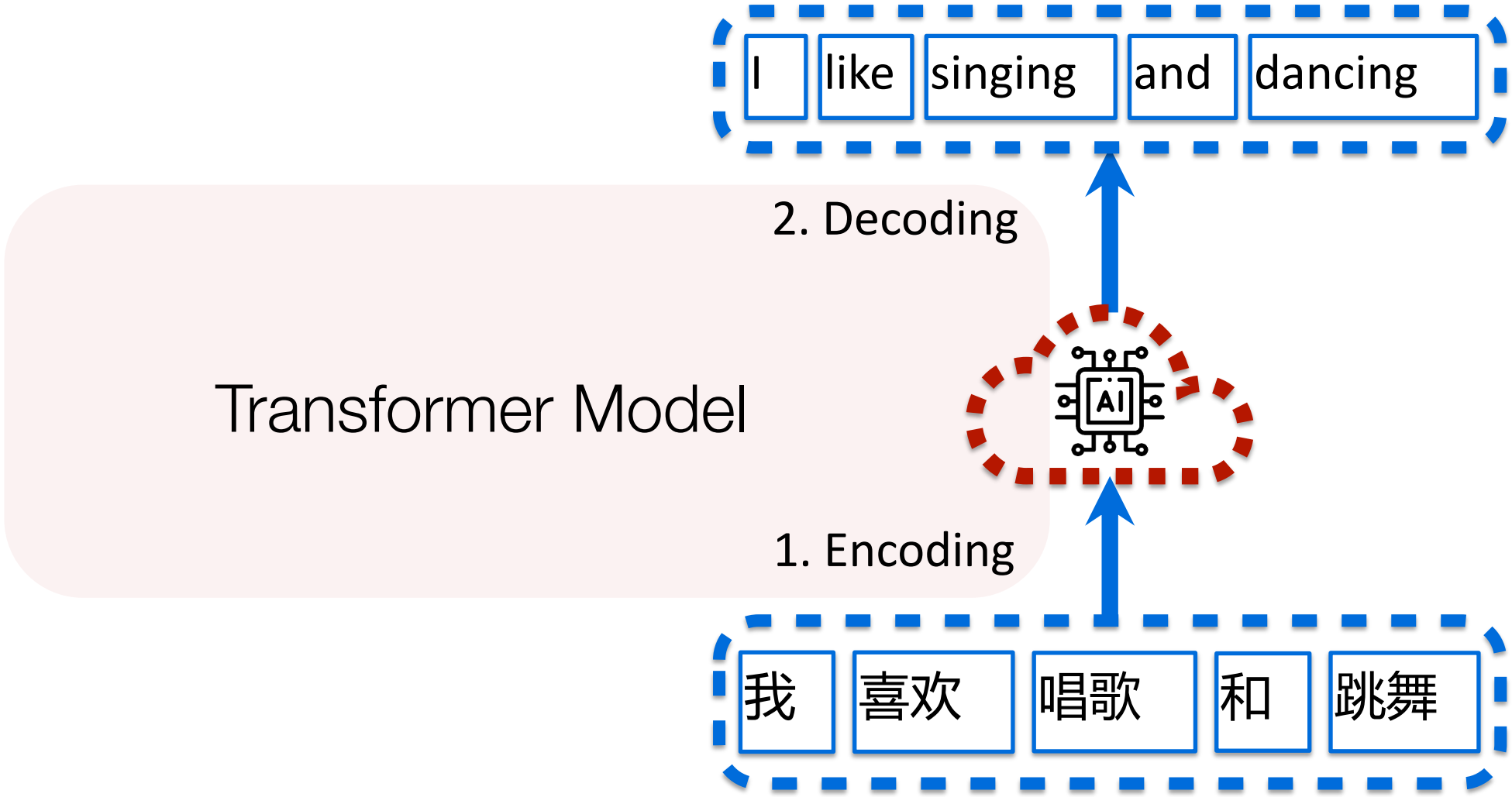
Erik Brynjolfsson,[a] Xiang Hui,[b] Meng Liu[b]

[a]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [b]Marketing, Olin School of Business, Washington University in St. Louis, St. Louis, Missouri 63130

World’s 7000 Languages



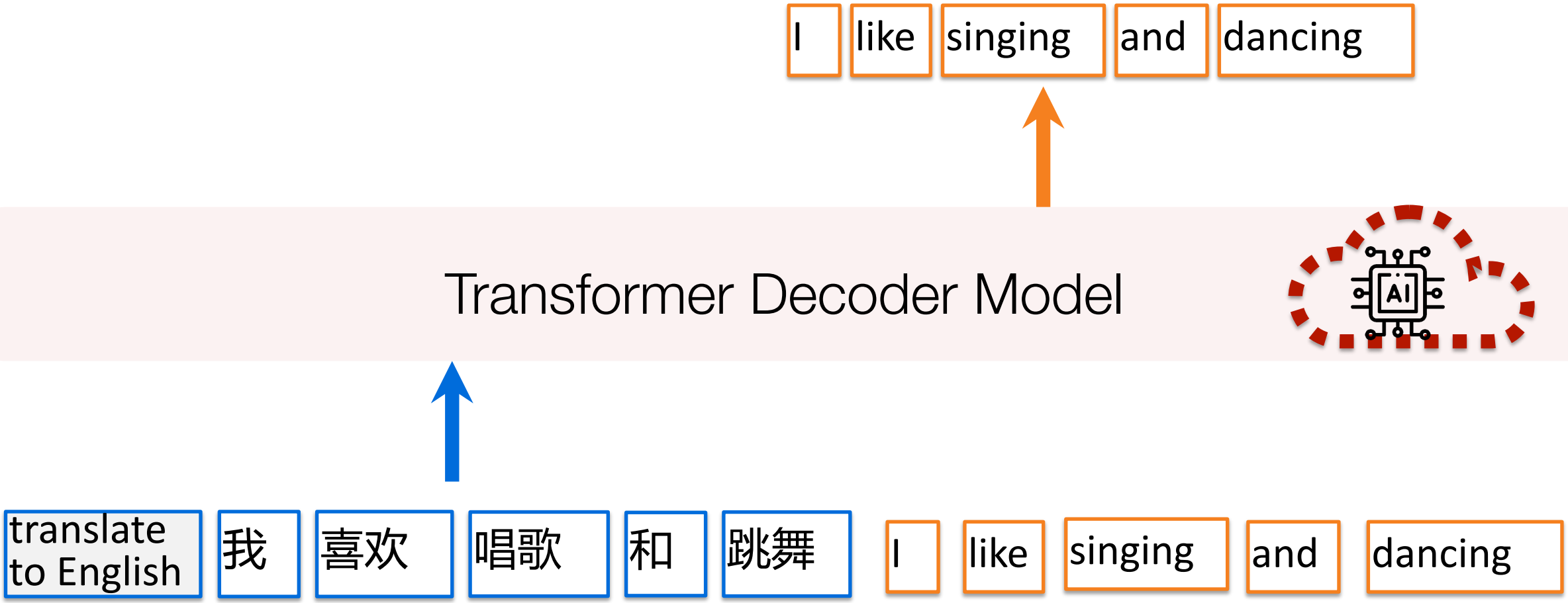
Neural Machine Translation



Attention is all you need. Vaswani et al 2017.

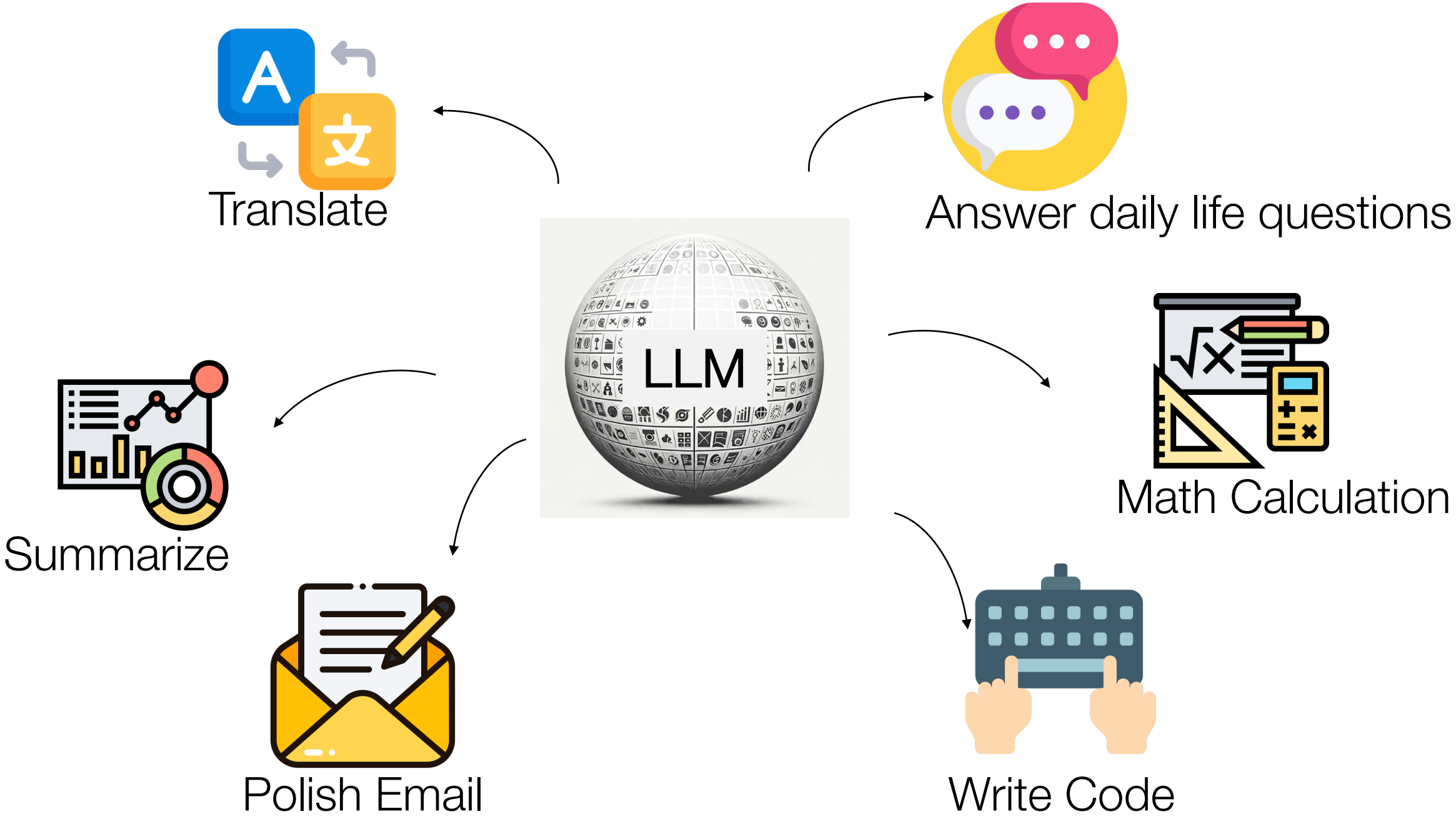
Sequence to sequence learning with Neural Networks. Sutskever et al 2014.

LLM for Translation



Attention is all you need. Vaswani et al 2017.

Sequence to sequence learning with Neural Networks. Sutskever et al 2014.


Translate
Answer daily life questions
LLM
Summarize
Math Calculation
Polish Email
Write Code

Outline

- The cross-lingual impact of vocabulary sharing in LLM
- LLaMAX: Scaling LLM to 100 languages
- LingoLLM: training-free method to enable LLM for endangered languages

Vocabulary

Word level

The most eager is Oregon which is enlisting 5 , 000 drivers in the country

Char level

T h e ␣ m o s t ␣ e a g e r ␣ i s ␣ O r e g o n ␣ …

Sub-word level

The most eager is O re go n which is en list ing 5 , 000 driver s in the country

Sub-word vocabulary is the dominant choice

Tokenizer – split text into basic units

Many words don’t map to one token: indivisible.

↓ tokenizer

Many words don't map to one token: indivisible.

[7085, 2456, 836, 470, 3975, 284, 530, 11241, 25, 773, 452, 12843, 13]

↓ embedding table lookup

2.3	-3.2	8.3	5.4	2.1	3.9	-8.9	3.8	3.9	3.3
4.5	5.9	4.5	7.1	1.0	5.3	5.0	3.1	0.7	5.0
…	…	…	…	…	…	…	…	…	…
3.8	1.2	3.8	9.0	9.3	3.1	4.2	0.8	9.2	5.8

Popular subword vocab: Byte-Pair-Encoding

- starting from chars
- repeatedly, merge most frequent pairs to form new tokens
- until reaching a fixed size.

raw word	freq.
cat	90
catch	50
rat	80
rattle	40

merge ('a', 't')

a
c
e
h
l
t
at

merge ('c', 'at')

a
c
e
h
l
t
at
cat

merge ('r', 'at')

a
c
e
h
l
t
at
cat
rat

merge ('cat', 'c')

a
c
e
h
l
t
at
cat
rat
catc

Neural Machine Translation of Rare Words with Subword Units. Sennrich et al. ACL 2016

Vocabulary Learning via Optimal Transport

- Entropy-regularized Optimal Transport

$$\min_{P \in \mathbb{R}^{m \times n}} \langle D, P \rangle - H(P)$$

subject to

$$\forall i \in Char, \sum_{j \in V_n} P_{i,j} = \hat{P}(i)$$

$$\forall j \in V_n, \left| \sum_{i \in Char} P_{i,j} - \hat{P}(j) \right| = \epsilon$$

- Sinkhorn's algorithm (from [Sinkhorn 1967])

Transportation matrix P

Char \ Tok	a	ab	bc
a	$P_{a,a}$	$P_{a,ab}$	$P_{a,bc}$
b	$P_{b,a}$	$P_{b,ab}$	$P_{b,bc}$
c	$P_{c,a}$	$P_{c,ab}$	$P_{c,bc}$

Cost matrix D

Char \ Tok	a	ab	bc
a	0	ln2	∞
b	∞	ln2	ln2
c	∞	∞	ln2

Xu, Zhou, Gan, Zheng, **Lei Li**. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

Vocabulary Sharing

English: television	**Spanish:** televisión
French: television	**Italian:** television
Dutch: televisie	**Portuguese:** televisão
Swedish: television	**Finnish:** televisio

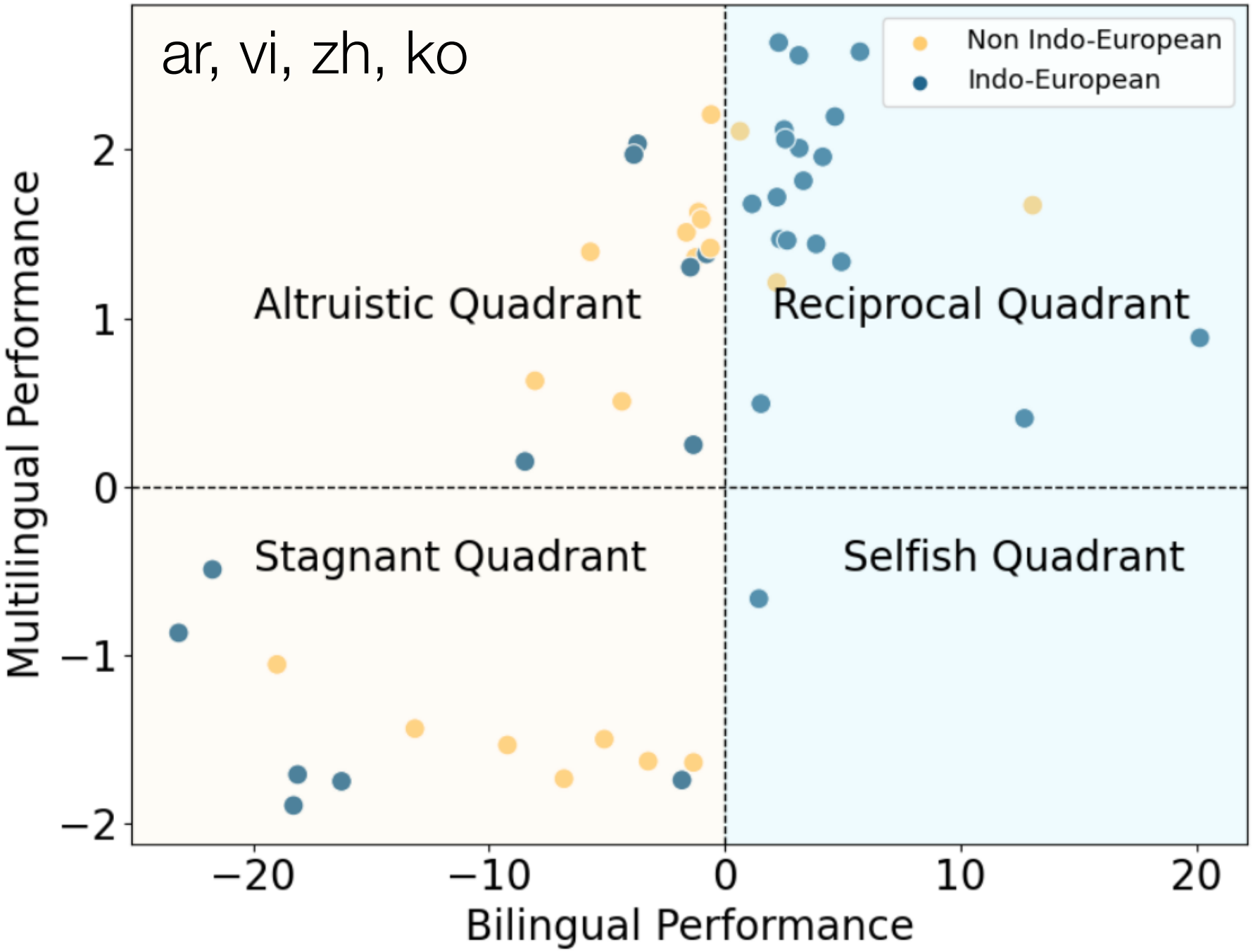
Embedding Finetuning for LLM

- Construct a small instruction-finetuning dataset using 10k bilingual parallel data
- Finetune LLaMA-7B
- Examine the translation performance of
  - The supervision bilingual direction (bilingual)
  - All other directions (multilingual)

Does embedding FT promote bilingual & multilingual translation performance?

Quadrant	Performance		Case Languages
	Bilingual	Multilingual	
Reciprocal	↑	↑	**cs, da, fr, de**
Altruistic	↓	↑	**ar, vi, zh, ko**
Stagnant	↓	↓	**Km, lo, gu, te**
Selfish	↑	↓	**hi**

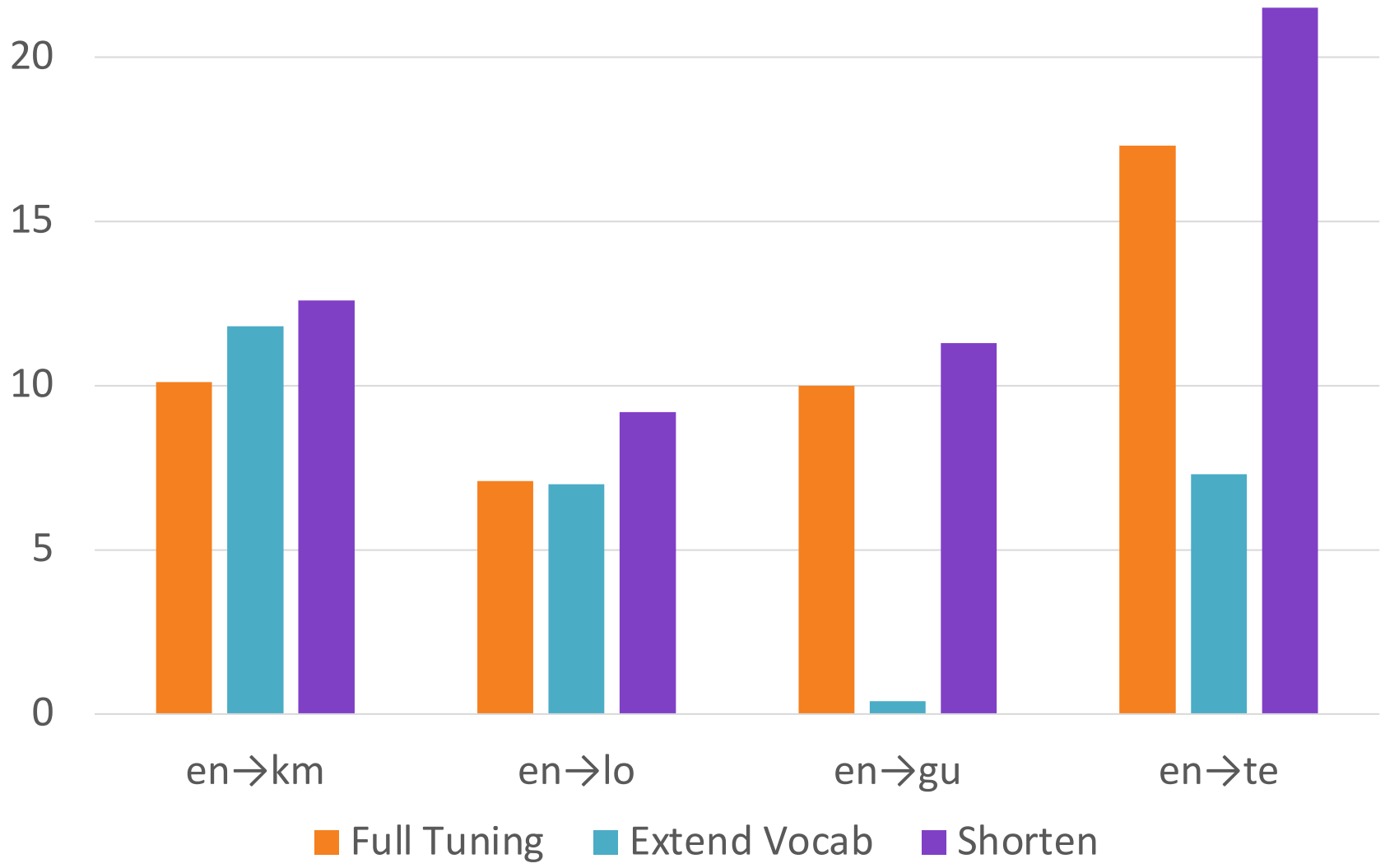
Fine-tuning on bilingual data does not always bring benefits to supervised direction!



Stagnant Quadrant – Over-tokenization

- Byte-BPE (BBPE) produces longer byte level token sequence than the number of characters
- 饕 [tāo] (gluttonous) ➔ three tokens [227, 234, 260]
- Implication for improvement:
  - shortening: remove the common prefix 227

Stagnant Quadrant: expanding vocab 👎 shortening 👍



Outline

- The cross-lingual impact of vocabulary sharing in LLM
- LLaMAX: Scaling LLM to 100 languages
- LingoLLM: training-free method to enable LLM for endangered languages

The quest of ~~multilingual~~ massive-lingual LLM

Release Date	Model	Base Model	Language	Model size	Affiliation
2024.02.12	Aya-101	mT5	101	13B	Cohere
2024.02.27	TowerLLM	LLaMA2	10	7/13B	Unbabel
2024.05.22	Aya-23	Command R	23	35B	Cohere
2024.06.24	Mistral Large 2	-	12	123B	Mistral AI
2024.07.08	LLaMAX	LLAMA3	101	7B	Shanghai AI Lab
2024.07.11	SeaLLM-v2.5	Gemma2	10	7B	DAMO, Alibaba
2024.07.31	LLaMA3.1	-	36	8/70/405B	Meta
2024.09.18	Qwen2.5	-	30	7/14/32/72B	Qwen, Alibaba
2024.09.26	EMMA500	LLaMA2	546	7B	University of Helsinki
2024.10.04	X-ALMA	LLaMA2	50	13B	Microsoft

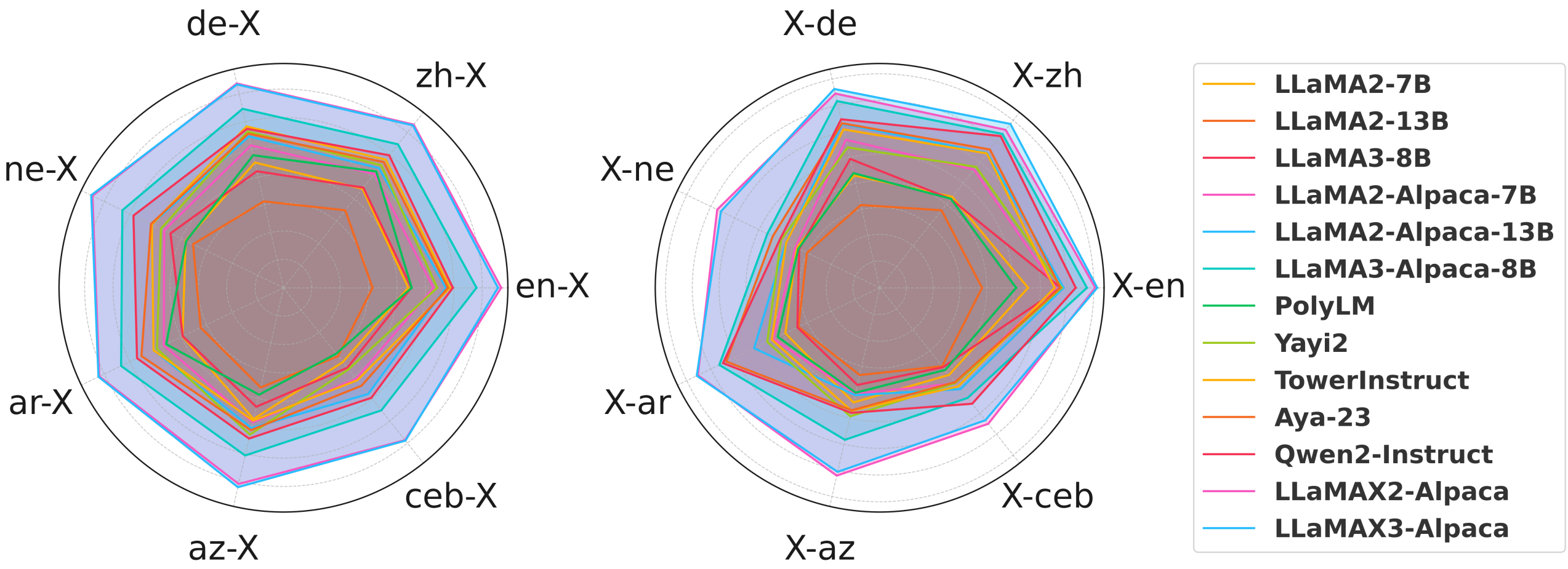
LLaMAX: continual pre-training + instruction fine-tuning

- Combine both parallel (102) and monolingual (94) data

- Data Augmentation by Random Aligned Substitution (RAS)

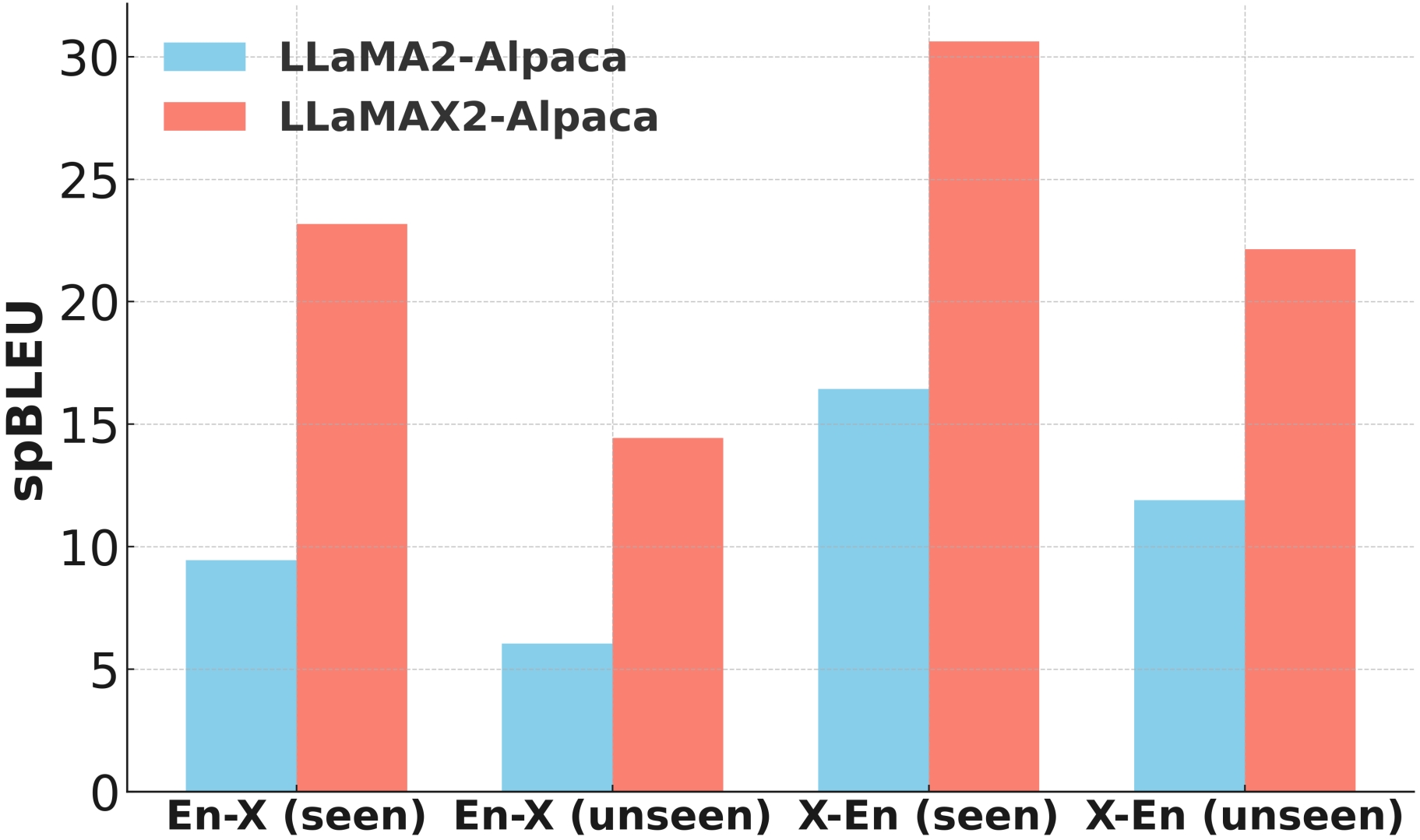
Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, Fei Yuan. LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages. EMNLP 2024.

LLaMAX achieves the best overall translation for 6<->101 langs



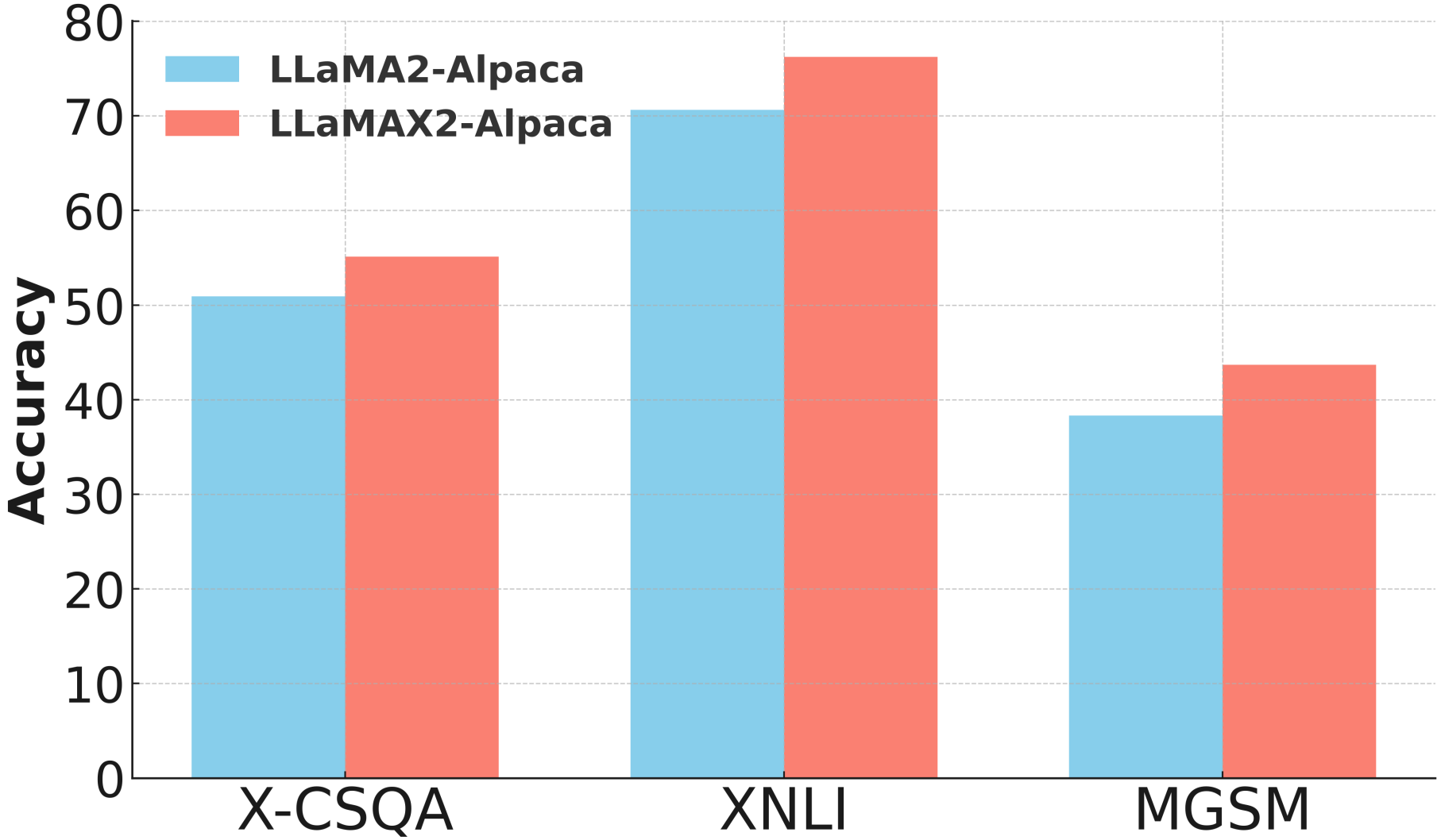
Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, Fei Yuan. LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages. EMNLP 2024.

LLaMAX improves translation for unseen languages



Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, Fei Yuan. LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages. EMNLP 2024.

LLaMAX is a better foundation model: retains and performs well on other multilingual tasks



Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, Fei Yuan. LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages. EMNLP 2024.

Outline

- The cross-lingual impact of vocabulary sharing in LLM
- LLaMAX: Scaling LLM to 100 languages
- LingoLLM: training-free method to enable LLM for endangered languages

LLMs cannot directly process endangered languages.

Translate this Manchu sentence into English:
bi yali qolame bahanarakv.

I still cannot move forward. ❌

I cannot stir-fry meat. ✅

Motivation: Using Linguistic Description in LLM

- **95%** of the world's 7000~ languages don't have enough data for training LLMs

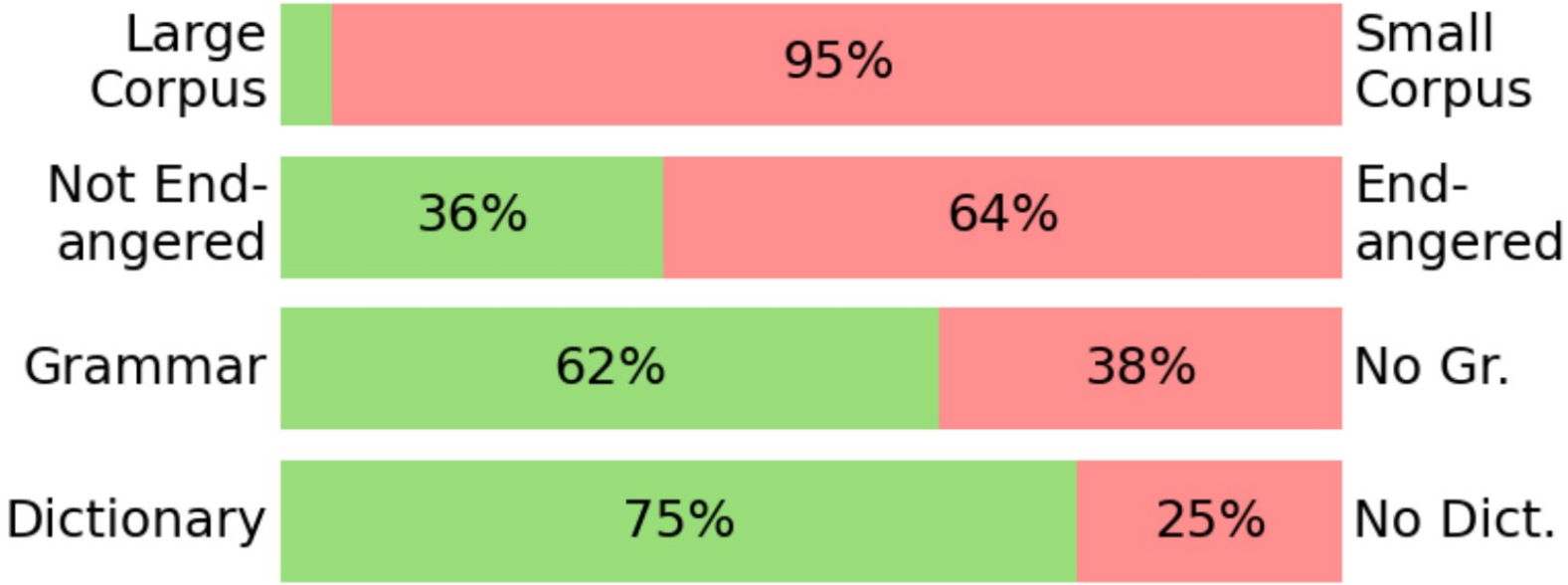
- Most have a grammar book (60%) or dictionary (75%).

Mandarin

English

Spanish

French

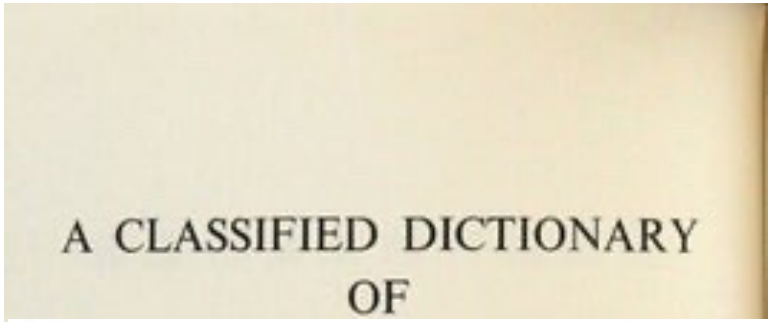


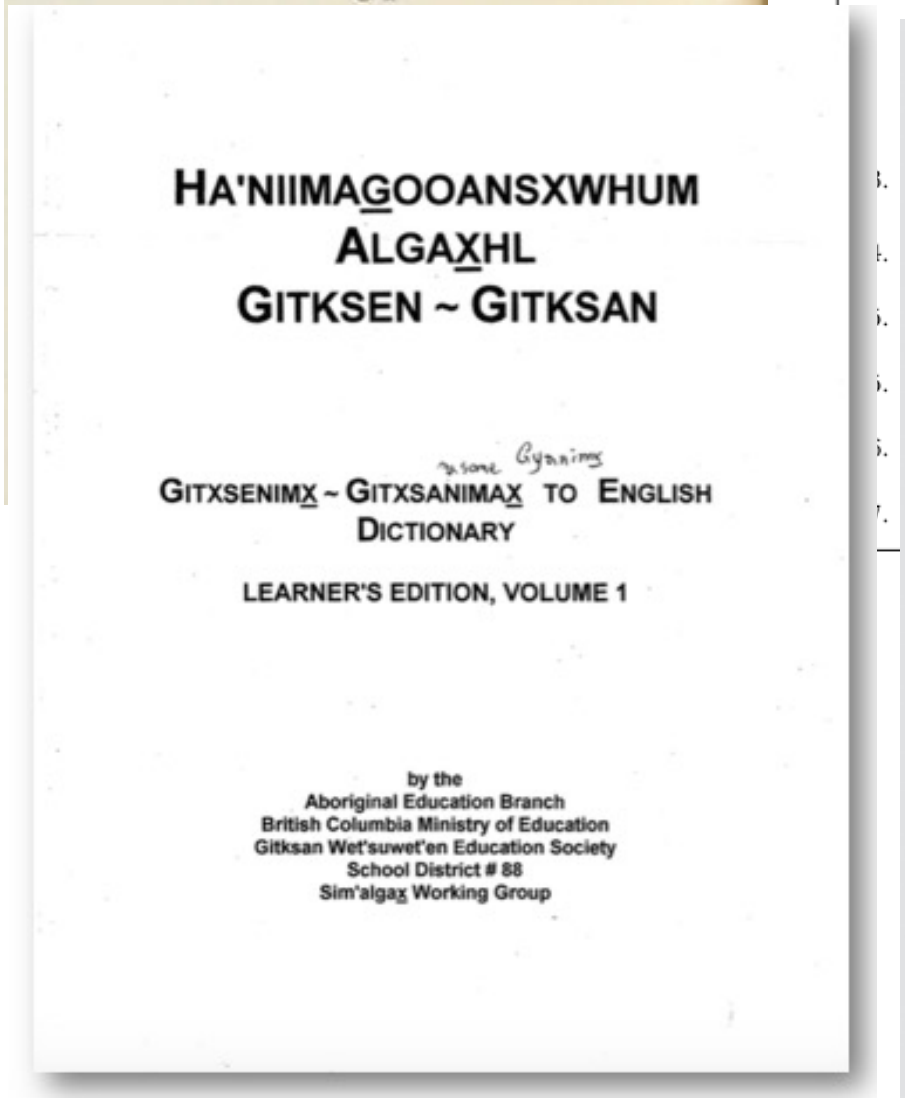
Cherokee

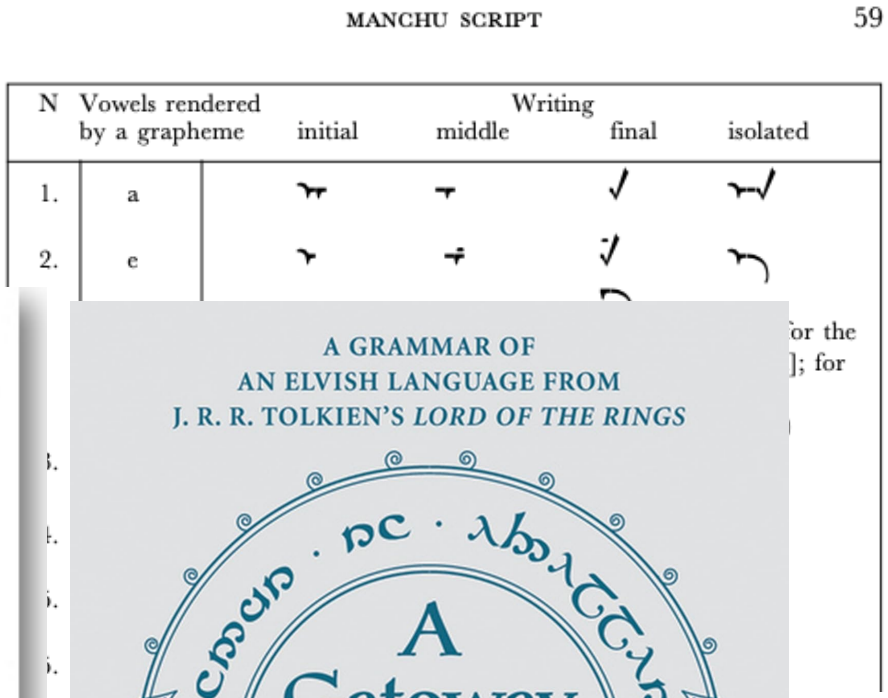
Gaelic

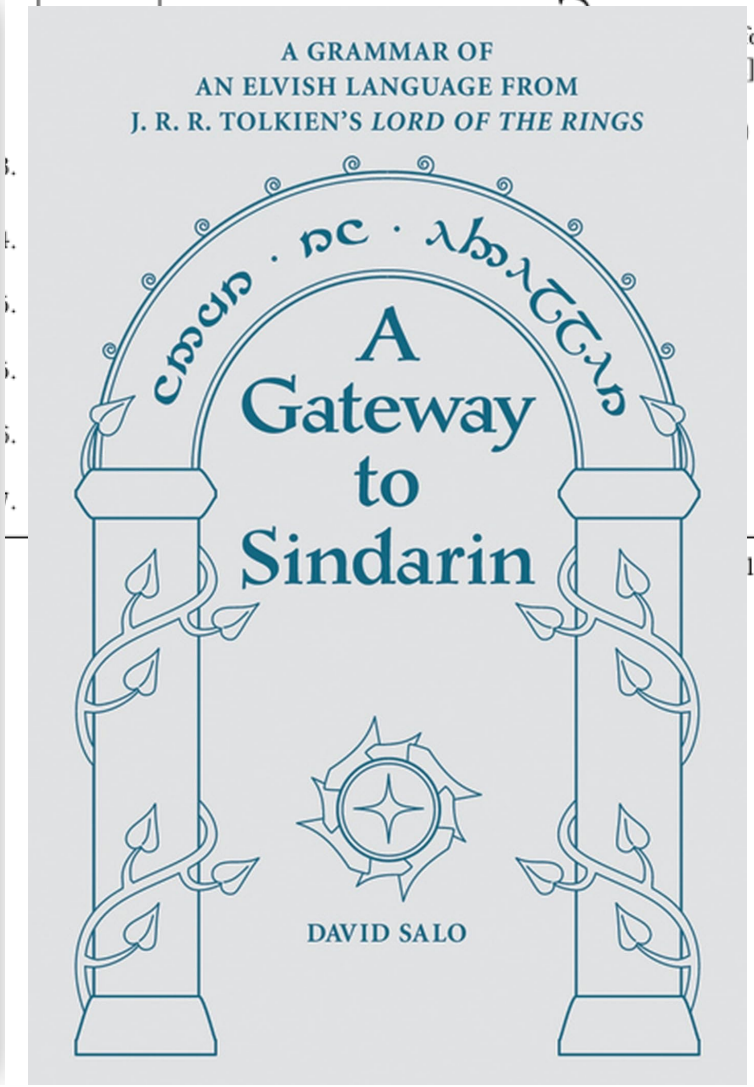
Wenzhou Wu

Tusom









Quenya

Late Period (1950-1973)

sindarin

Q. *noun.* Grey-elven

Element in

- Q. hwesta sindarinwa "Grey-elven hw"
  ✧ LotR/1123

Elements

Word	Gloss
Sinda	"Grey-elf"
-rin	"-ian, racial-adjective, language"

[LBI/Sindarin; Let/176; Let/219; LotR/1123; LotR/1127; LotRI/Sindarin; LRI/Sindar;

sindë

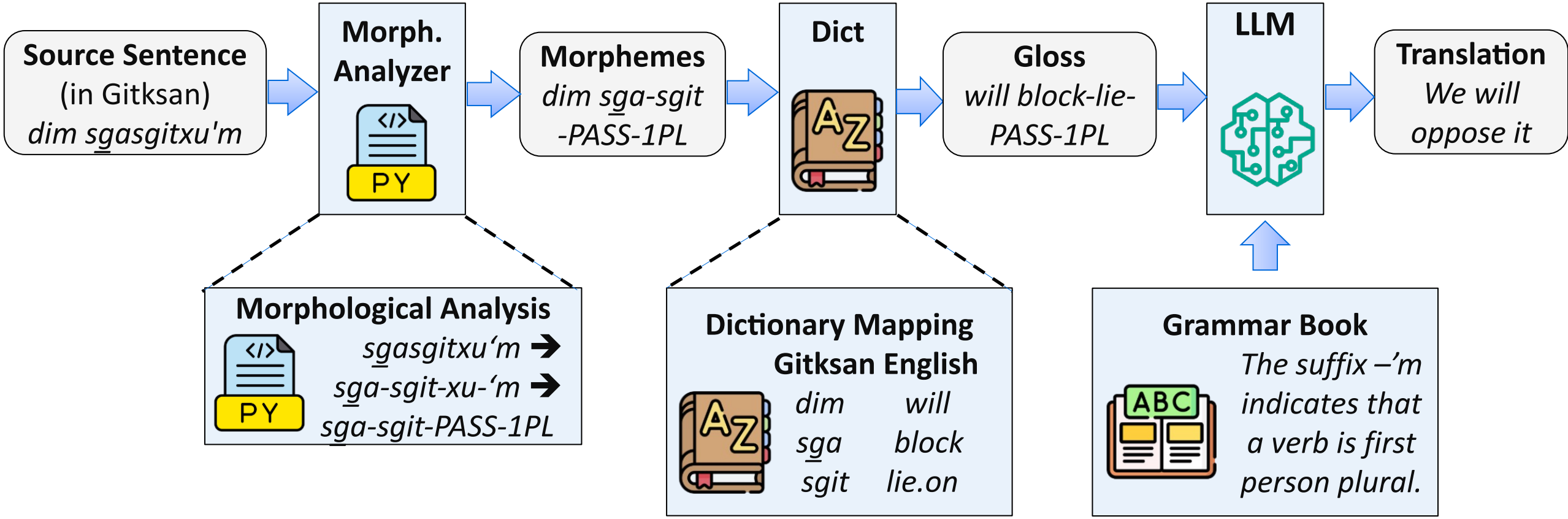
Q. grey, pale or silvery grey

sindë (þ) *adj.* "grey, pale or silvery grey" (the Vanyarin dialect preserves the older form **þindë**) *(WJ:384, THIN; in SA:thin(d) the form given is* ***sinda****, cf. also* ***sindanóriello*** *"from a grey country" in Namárië.* ***Sindë*** *and* ***sinda_*** are apparently variants of the sam word.) _Stem **sindi-**, given the primitive form ¤*thindi*; cf. **Sindicollo** (q.v.)

[Quettaparma Quenyallo] Group: Quettaparma Quenyallo. Published 11 years ago by Ardalambion (Helge Fauskanger).

LingoLLM

Insight: Make LLMs translate like human language learners.



Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

LingoLLM Step 1: Morphological Analysis

- Turn words into morphemes:
  - easier to find in dictionaries; we know their roles in a sentence.

- An example in English: Cats got your tongue.

Word	Morphemes	
cats	Cat+Plural	
got	get+Past	
your	2nd.Person.Singular+Possession	
tongue	Tongue+Singular	

Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

LingoLLM Step 2: Dictionary Matching

- Find the closest match in the dictionary (not always exact)
- An example in English to Chinese: **Cats got your tongue.**

Word	Morphemes	Mapped Morphemes
cats	Cat+Plural	猫+Plural
got	get+Past	拿到+Past
your	2nd.Person.Singular+Possession	你+Possession
tongue	Tongue+Singular	舌头+Singular

Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

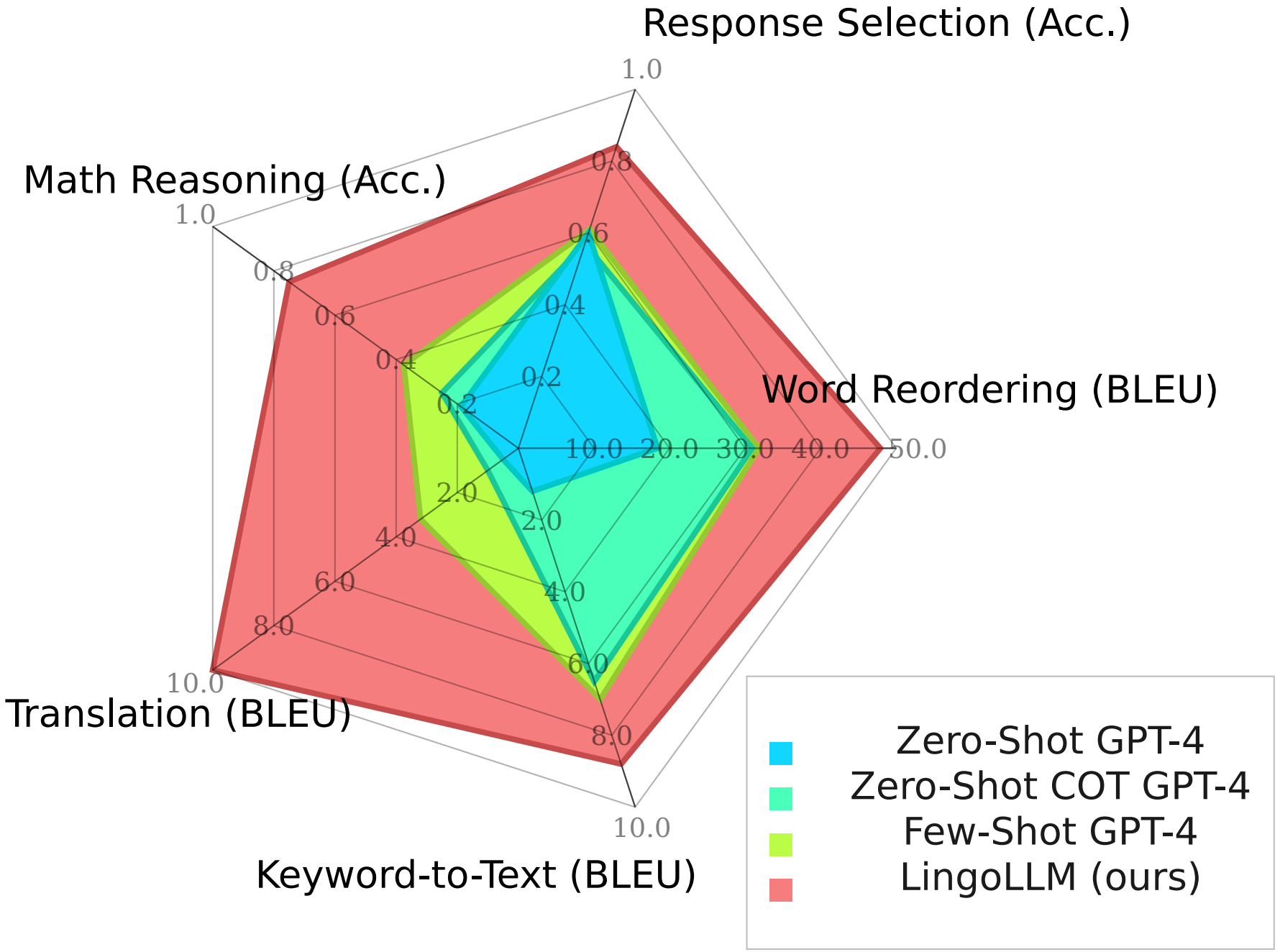
LingoLLM Step 3: LLM Translation

This is a grammar book for Manchu.
Manchu has a subject-object-verb word order.
Translate the following sentence from Manchu to English:
bi yali qolame bahanarakv.
Here's the word by word translation of the words:
bi – I; yali – meat; qolame – stir-fry.PRESENT;
bahanarakv - cannot.PAST.IMPERFECT;
Translate the sentence into English.

Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

LingoLLM significantly improves NLP on endangered languages

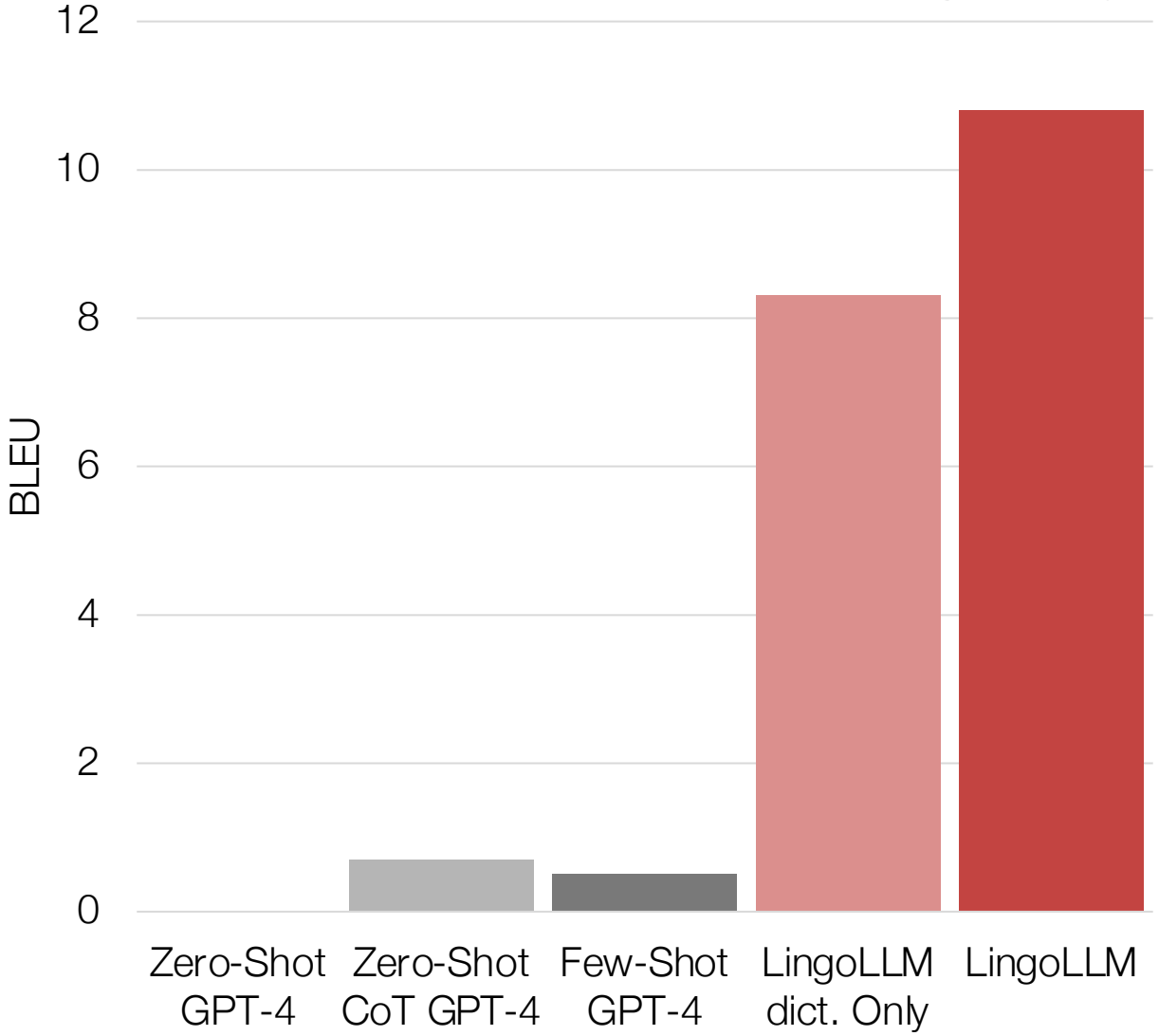
5 tasks,
8 languages

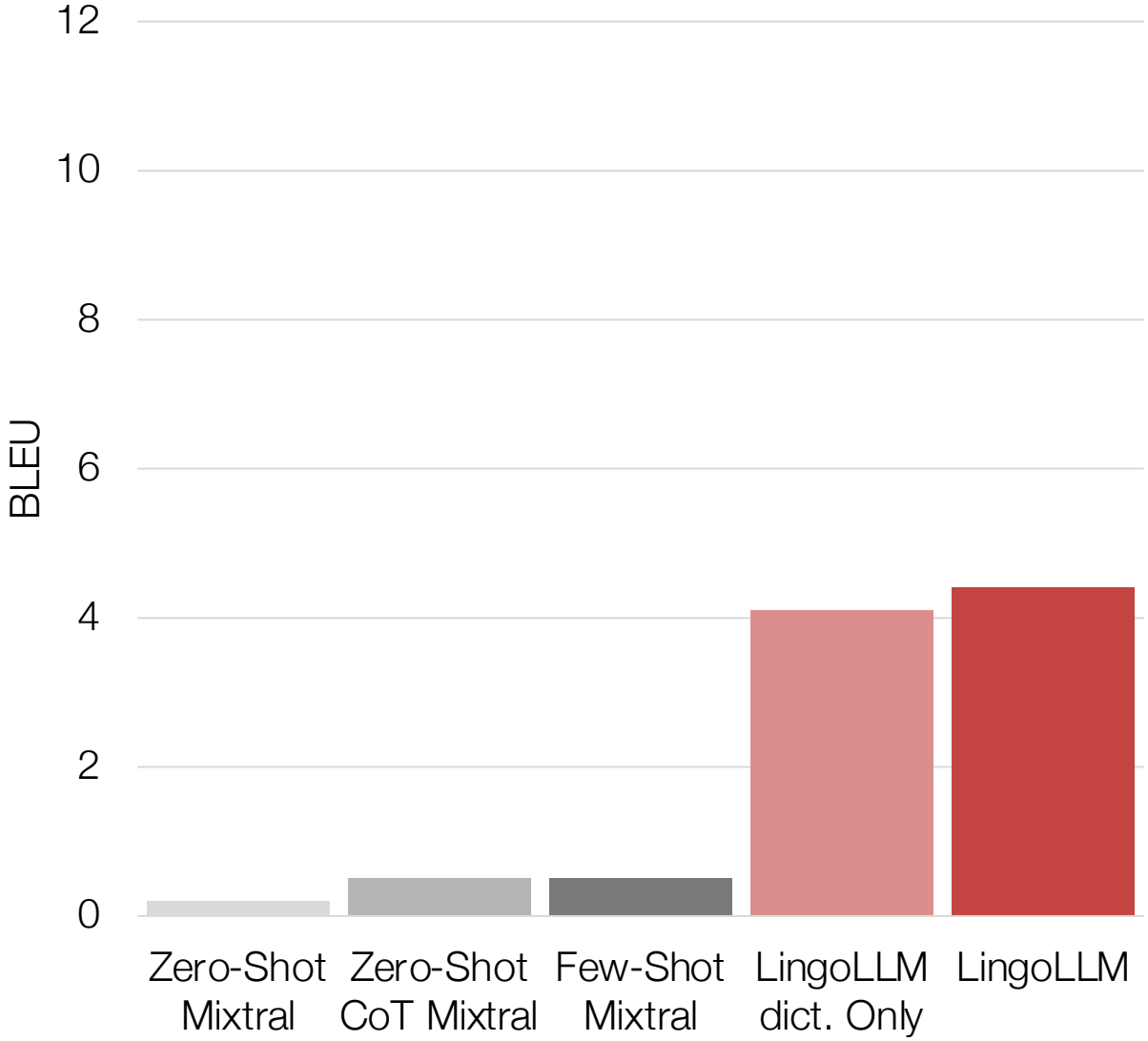


Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

LingoLLM elevates LLM from no-translation to decent translation

Manchu, Gitksan, Natugu, Arapaho, Uspanteko, Tsez, Bribri, Wolof





Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

LingoLLM Translation Case - Manchu

	Manchu
Input	suweni geren xusai dorgi de nikan i niyalma udu qoohiyan i niyalma udu
Reference	Among your many students, how many are Chinese and how many are Korean?
GPT4 - fewshot	Every person in the military and every person in the common people must have courage
LingoLLM	How many Chinese people and how many Koreans are there among your numerous students?

Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

LingoLLM Translation Case - Gitksan

	Gitksan
Input	Way ts’ax wildiihl hehl Gitwinhlguu’l ii needii hasakdiitehl reserve. “Needii hasaga’m dim dip suwii gi’namhl laxyibi’m,"dihiida.
Reference	And now even though the people of Kitwancool said they did not want the little reserve; “We don’t want to give away our land," they said.
GPT4 - fewshot	He said, "I will stay here in Gitanyow, and you will go to the reserve. ’You will learn to speak English well there,’ he told me."
LingoLLM	"Although it seems that the people of Kitwancool don’t want the reserve, ’We do not wish to give away our land,’” they said.

Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

LingoLLM Translation Case - Arapaho

	Arapaho
Input	nihcihcee3ciiteit niiyou nuh'uuno heenees3i'okuutooni'
Reference	He inadvertently walked in where people were sitting.
GPT4 - fewshot	I'm going to work for you tomorrow.
LingoLLM	Someone accidentally entered this room where people sit.

Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

Can LingoLLM solve a math problem in an endangered language?

Example Problem (Manchu): Mari qi Jon juwe (2) se ajigesi, Jon qi Jeisa sunja (5) se amba. aika Jeisa 20 se oqi, ere ilan (3) sarganjui i se be uheri aqaqi yagese ombi?

Example Problem (English): Mary is two years younger than Joan, who is five years older than Jessa. If Jessa is 20 years old, what is the sum of the ages of the three girls?

Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

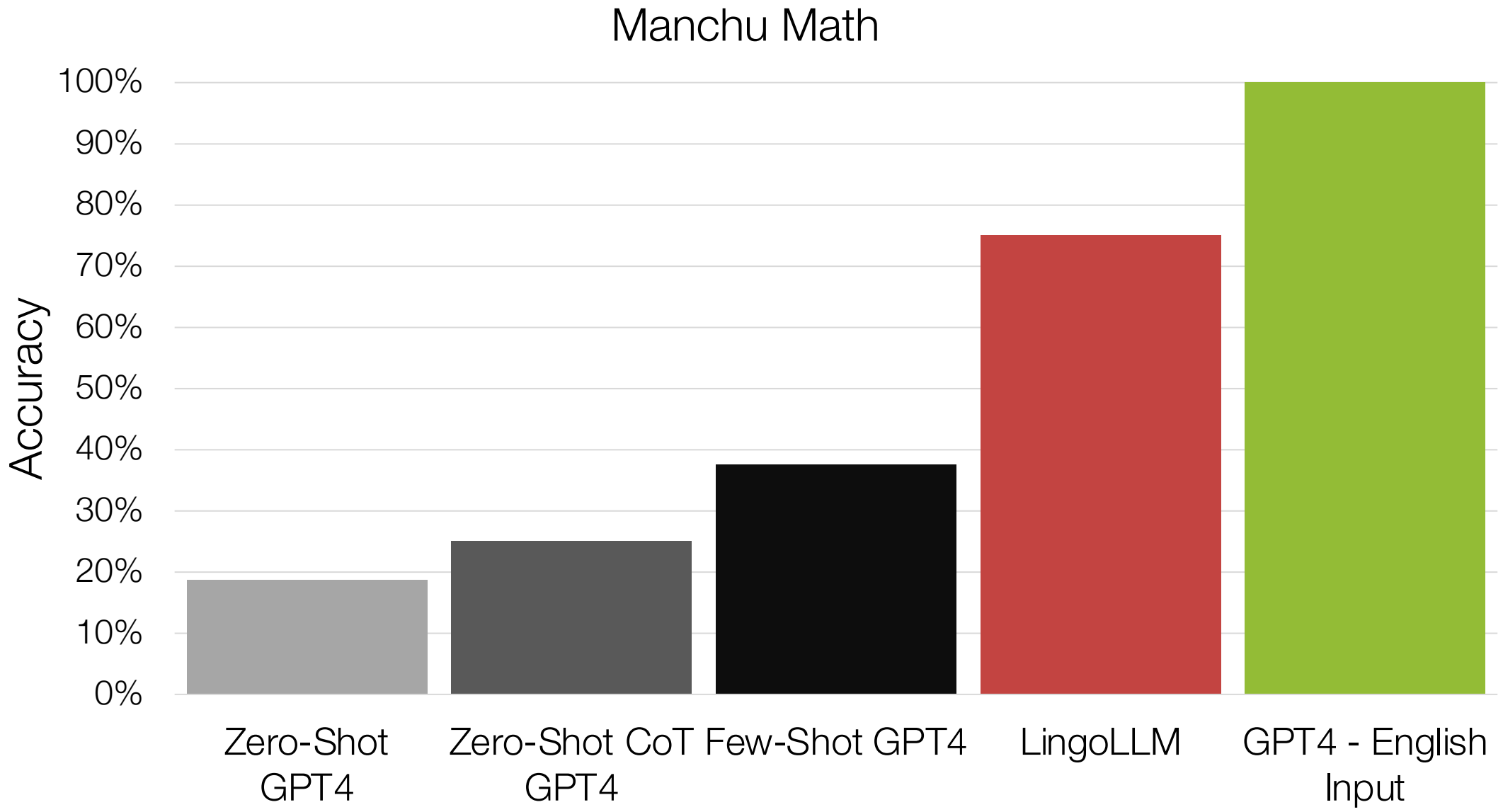
LingoLLM solves a math problem by translating it first

Original Problem: Baldur gets water from a well. He gets 5 pails of water every morning and 6 pails of water every afternoon. If each pail contains 5 liters of water, how many liters of water does he get every day?

LingoLLM translation from Manchu: Balder, early in the morning, picks up water from the well. He takes five buckets in the evening, and six buckets in the morning. If one bucket equals five bowls, how many bowls of water does he get in a day?

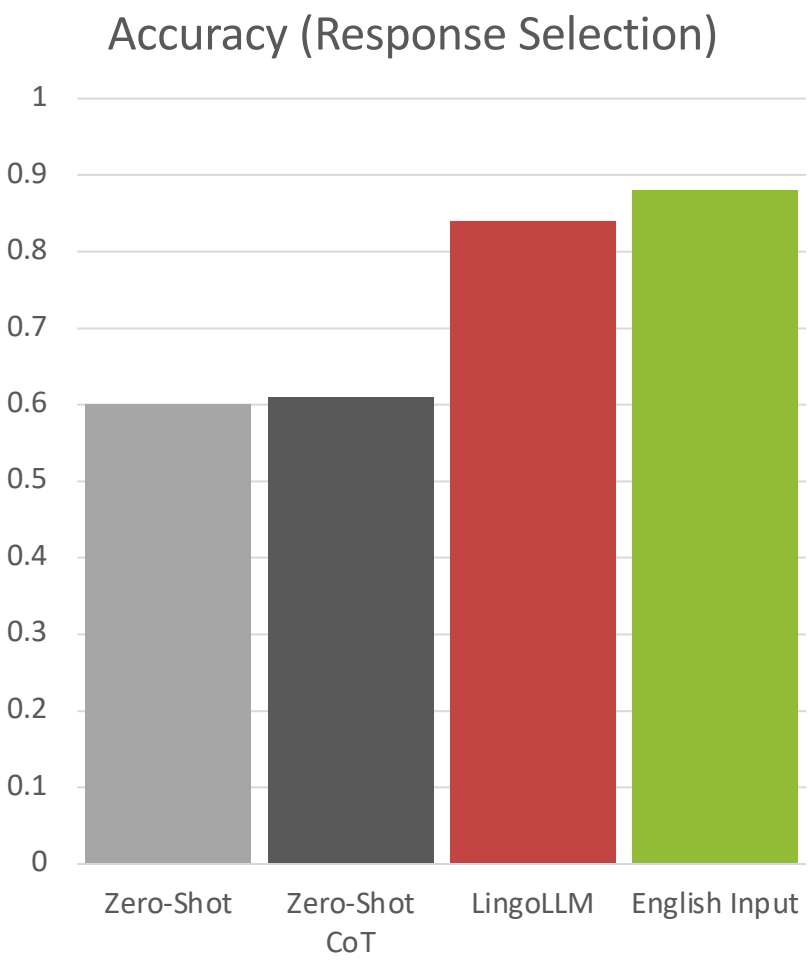
Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

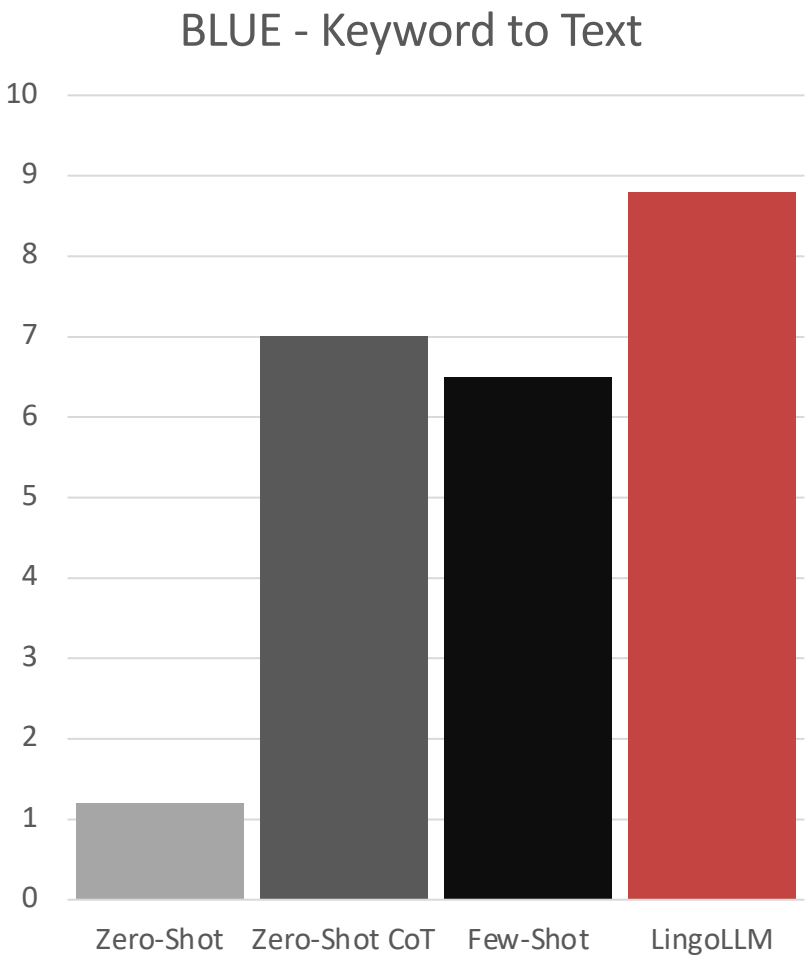
LingoLLM Manchu gets close to English in math

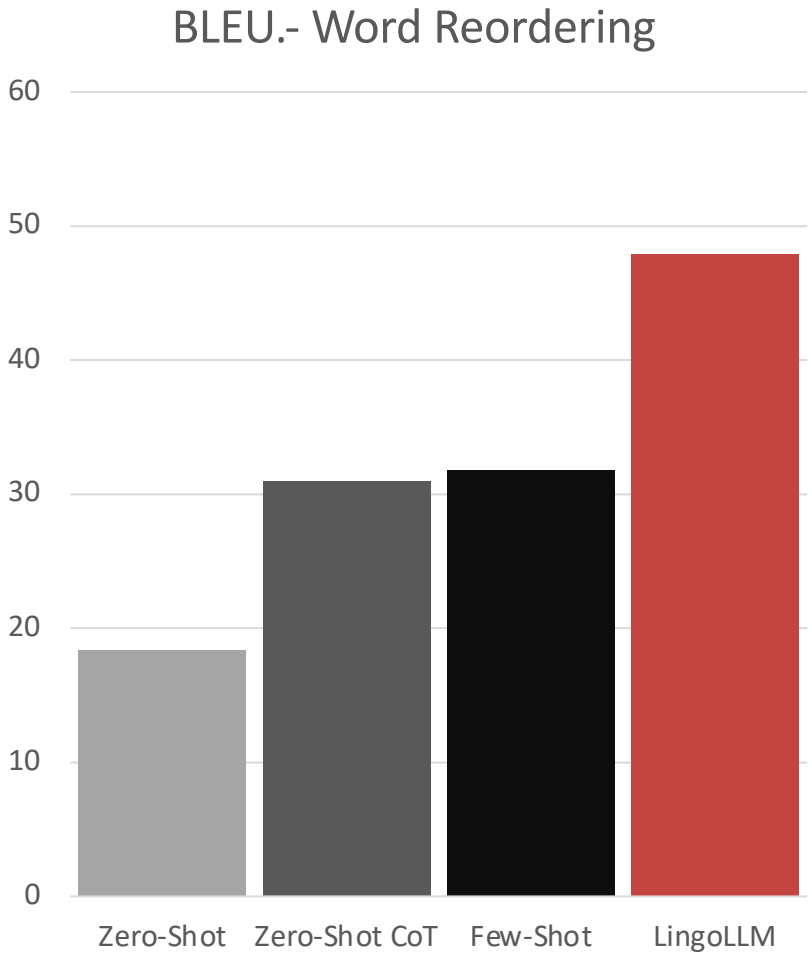


Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

LingoLLM performs well in multiple tasks and languages.







Zhang, Choi, Song, He, Wang, Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. ACL-Findings 2024.

Summary

- Vocabulary sharing leads to different four-quadrant impact
  - Altruistic: bilingual fine-tuning improves other language
  - Stagnant: shortening helps
- LLaMAX: Scaling LLM to 100 languages
  - do not expand vocab!
  - combining both bilingual and monolingual
  - data augmentation
- LingoLLM: using morphological analyzers, dictionaries, and grammar books to enable LLM for endangered languages

Multilingual Translation @ Li-Lab

Vocabulary Construction

Neurips 19
VOLT, ACL 21a
LLaMA vocab, ACL 24

Training

mRASP, EMNLP20
mRASP2, ACL21b
LaSS, ACL 21c
CIAT, EMNLP 21a
REDER, NeurIPS 21
MGNMT, ICLR 20
Prune-tune, AAAI 21
LegoMT, ACL 23

Evaluation

SEScore, EMNLP 22
SEScore2, ACL 23
InstructScore, EMNLP23
Translate-Canvas, EMNLP 24

Deploy

KSTER, EMNLP 21c

Serving/ Inference

CapsNMT, EMNLP 19
GLAT, ACL 21e
latent-GLAT, ACL 22
REDER, Neurips 21
LPDS, AAAI 22
switch-GLAT, ICLR 22
ICML 22

Speech Translation

WACO, ACL23
ConST, NAACL 22
MoSST, ACL 22a
STEMM, ACL 22b,
Chimera, ACL 21d,
LUT, AAAI 21b,
COSTT, AAAI 21c
XSTNet, Interspeech 21
NeurST, ACL 21

LLM for MT

Graformer, EMNLP 21b
CTNMT, AAAI 20
LLM-trans-benchmarking, NAACL24
LLMRefine, NAACL 24
LingoLLM, ACL 24
LLaMAX, EMNLP 24

Acceleration

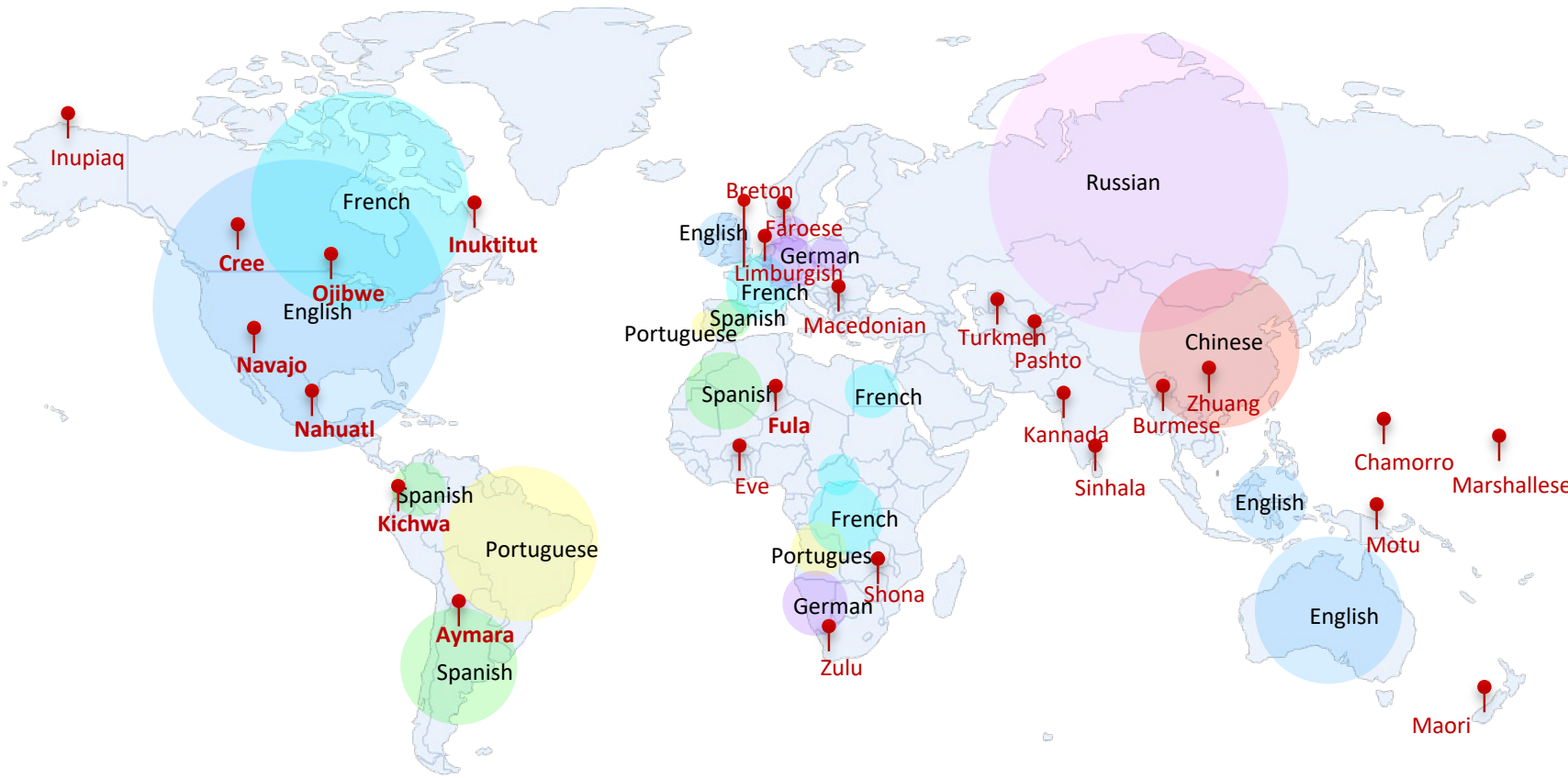
LightSeq, NAACL21
LightSeq2, SC22

Human Interaction

CAMIT, IJCAI 19

Crossing Barriers for 1000 Languages

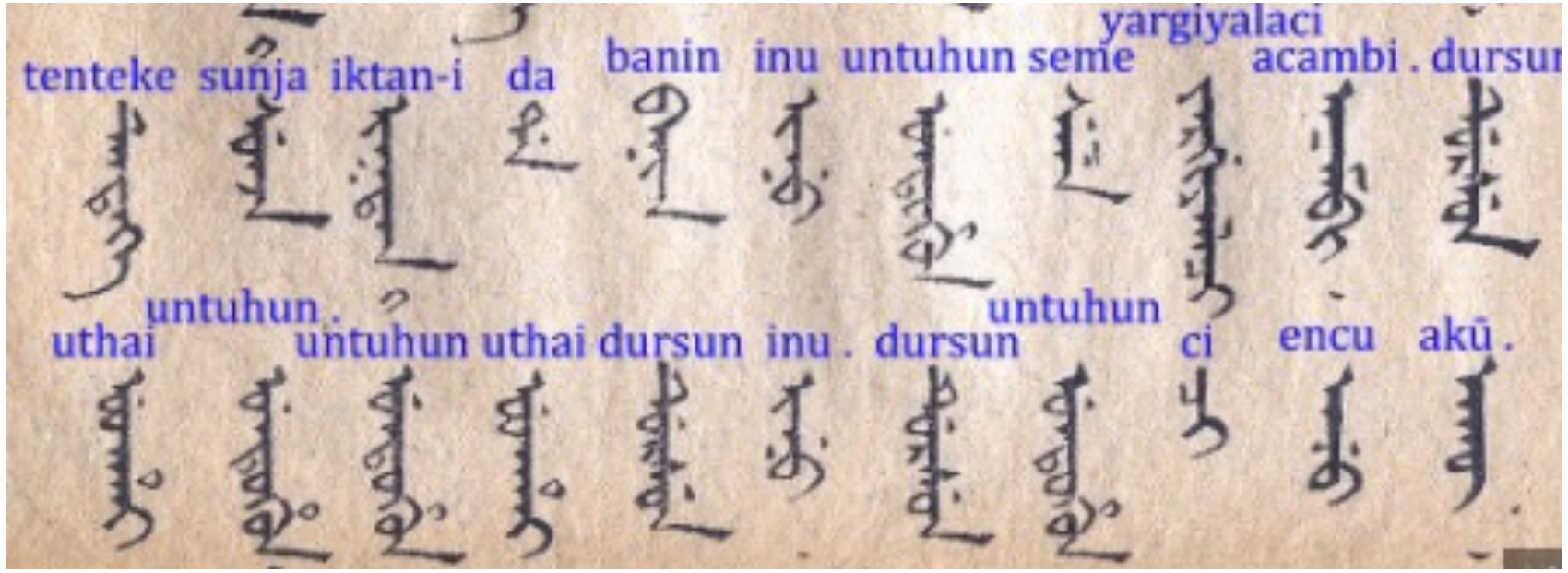
- 1: Democratizing MT for extremely-low resource languages
  <10k parallel sentences
- 2: Low-latency Streaming Speech Translation
  <100hrs speech-text data
- 3: Efficient, low-cost Translation



understudied

Challenges

- **Script:** Endangered languages may have rare scripts and orthography.



Challenges

- **Script:** Endangered languages may have rare scripts and orthography.
- **Speech:** Endangered languages may only be spoken and not written. It's more useful if we process speech.
- **OCR:** Many dictionaries and grammar books are not well-digitized.
- **Agentic:** LingoLLM may perform better if it can locate context more autonomously with given tools, instead of following a fixed workflow.

Towards Scaling Large Language Models to 1000 Languages

VOLT

mRASP

WACO

LEGOMT

ConST

LLaMAX

LightSeq

https://leililab.github.io