LLM Multilinguality

Tomasz Limisiewicz

25 April 2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Multilingual LLMs

Why do we train Multilingual LLMs?

Multilingual LLMs

Why do we train Multilingual LLMs?

💡 Accessibility to technology for speakers across the Globe 🌐

💡 Efficiency: more sustainable than training a model for each language

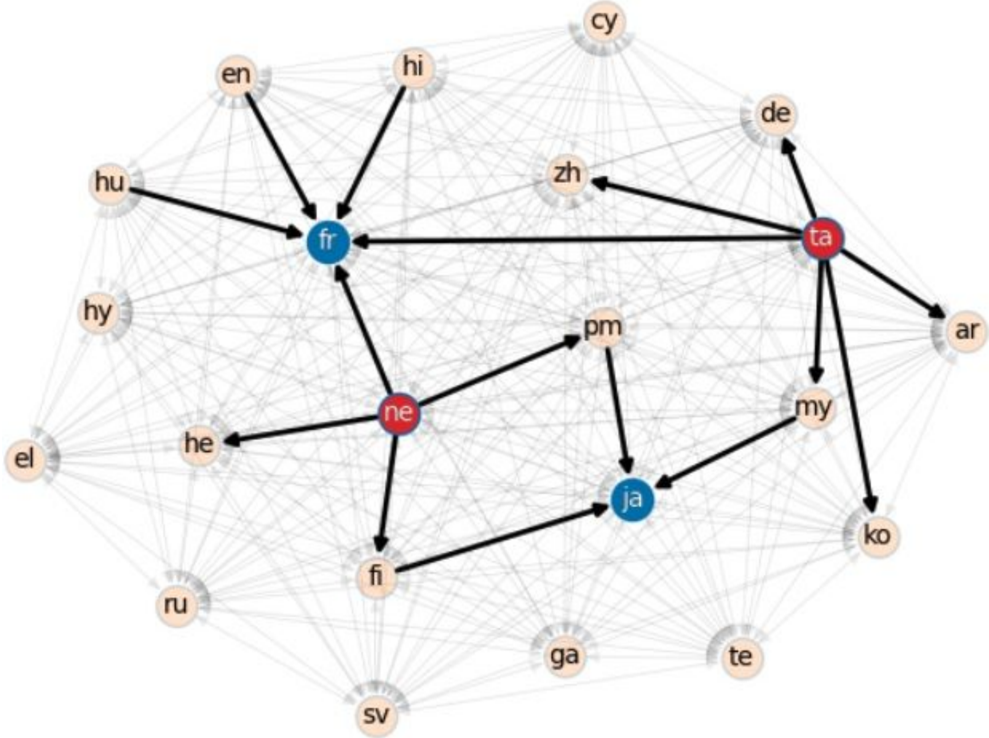
💡 Cross-lingual transfer learning

Multilingual LLMs: Cross-Lingual Transfer

source: https://arxiv.org/pdf/2205.04086

Model share knowledge of multiple languages

Fine-tuning or instructing the model for a task in one language enables solving it for another one



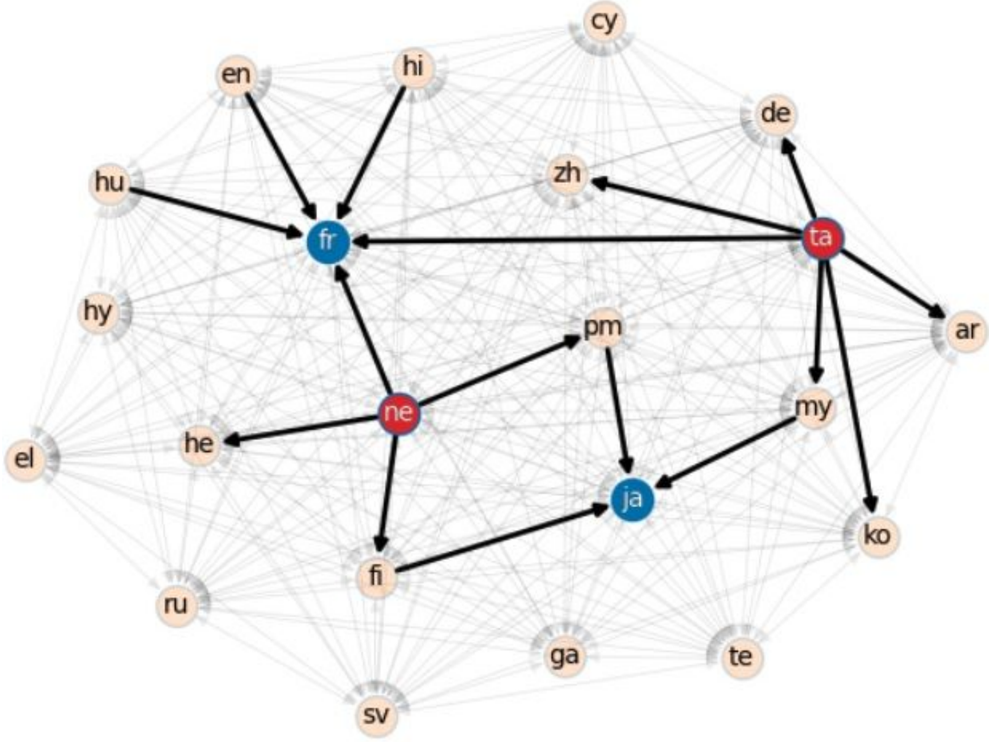
Multilingual LLMs: Cross-Lingual Transfer

source: https://arxiv.org/pdf/2205.04086

Model share knowledge of multiple languages.

Fine-tuning or instructing the model for a task in one language enables solving it for another one.

Not yet fully understood, but we have some clues!



Multilingual LLMs: Cross-Lingual Transfer

How to pick the right model for my* language?

* *For English speakers: my friend's language*

Multilingual LLMs: Cross-Lingual Transfer

How to pick the right model for my* language?

💡 Check if model claims to support it

** For English speakers: my friend's language*

Multilingual LLMs: Cross-Lingual Transfer

How to pick the right model for my* language?

💡 Check if model claims to support it

💡 Check the training or fine-tuning data

💡 Check data for similar languages

** For English speakers: my friend's language*

Multilingual LLMs: Cross-Lingual Transfer

How to pick the right model for my* language?

💡 Check if model claims to support it

💡 Check the training or fine-tuning data

💡 Check data for similar languages

💡 Check tokenization

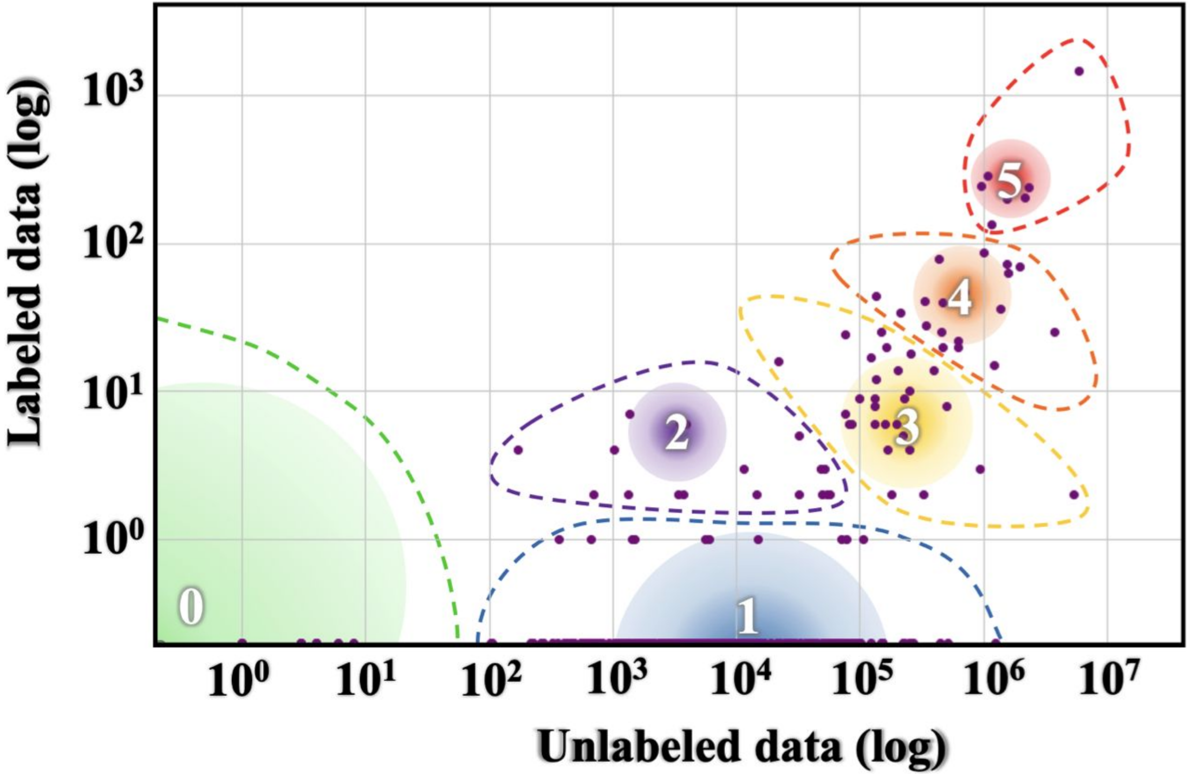
** For English speakers: my friend's language*

Let’s Talk About Data

Multilingual LLMs: Data

source: https://arxiv.org/pdf/2004.09095

Languages hugely differ
In data availability.



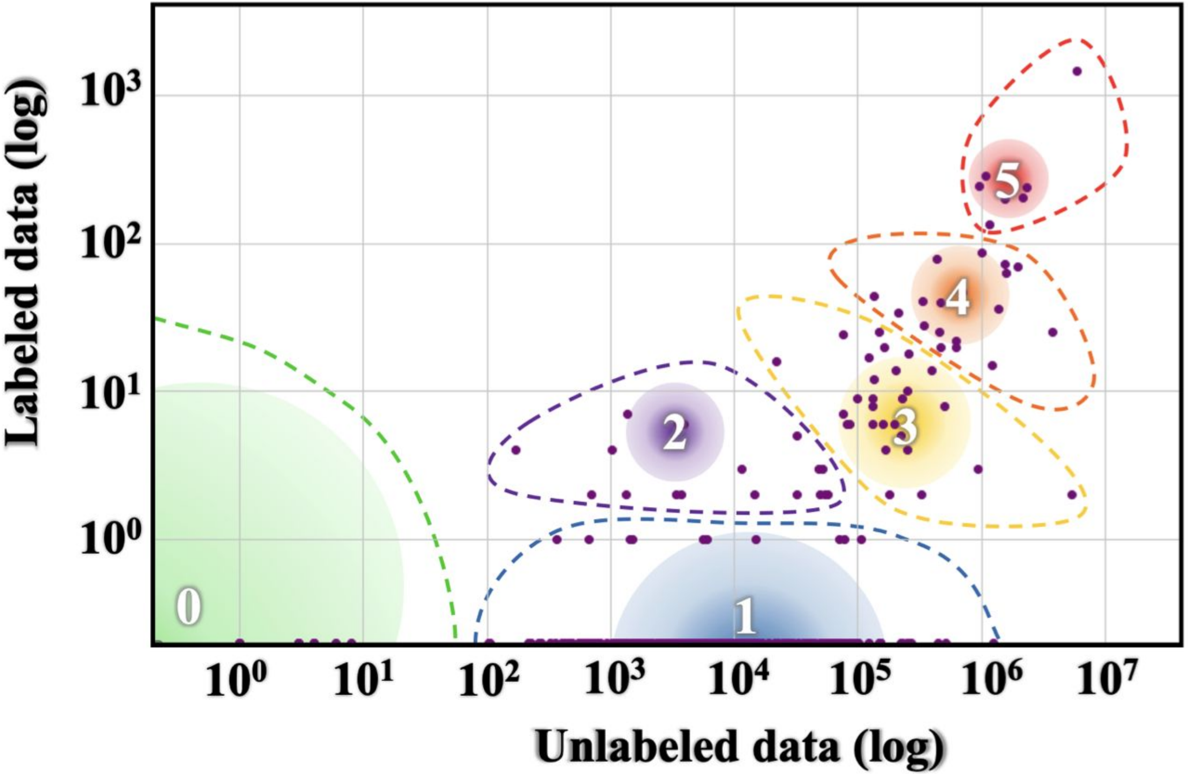
Multilingual LLMs: Data

source: https://arxiv.org/pdf/2004.09095

Languages hugely differ
In data availability.

Unlabeled data for
pre-training

Labeled data for
tuning and evaluation



Discussion time!

1. How resourceful is your language?

2. Do you think that it is underrepresented?

How Resourceful is Your Language?

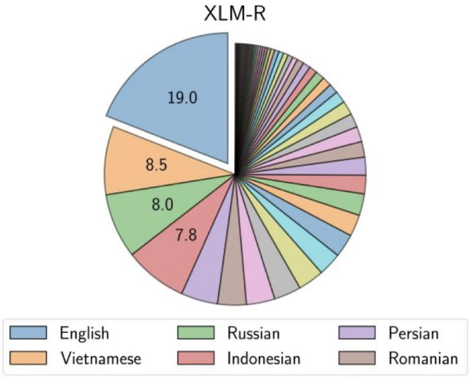
Group	Speakers		Languages
5	2.5 B		English, Spanish, German, French, Arabic, Mandarin
4	1.6 B		Russian, Hungarian, Vietnamese, Czech, Polish, Persian, Hindi
3	1.1 B		Indonesian, Ukrainian, Hebrew, Cebuano, Slovak
2	300 M		Irish, Maltese, Lao, Zulu, Amharic
1	1 B		Cherokee, Fijian, Greenlandic, Navajo, Macedonian
0	1 B		Dhalo, Warlpiri, Popoloca, Wallisian, Bora

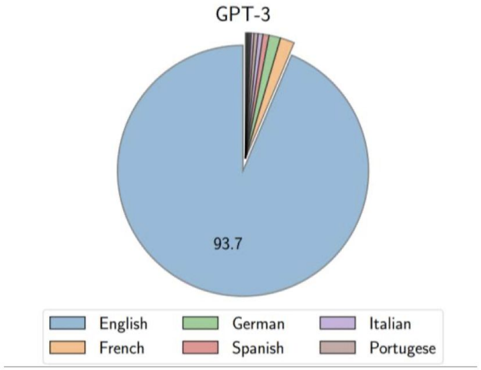
How Resourceful is Your Language?

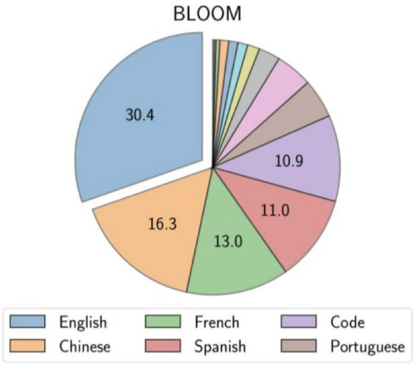
Group	Speakers		Languages
5	2.5 B		English, Spanish, German, French, Arabic, Mandarin
4	1.6 B		Russian, Hungarian, Vietnamese, Czech, Polish, Persian, Hindi
3	1.1 B		Indonesian, Ukrainian, Hebrew, Cebuano, Slovak
2	300 M		Irish, Maltese, Lao, Zulu, Amharic
1	1 B		Cherokee, Fijian, Greenlandic, Navajo, Macedonian
0	1 B		Dhalo, Warlpiri, Popoloca, Wallisian, Bora

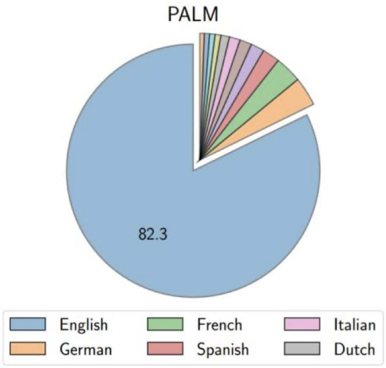
Multilingual LLMs: Data

How to balance data?





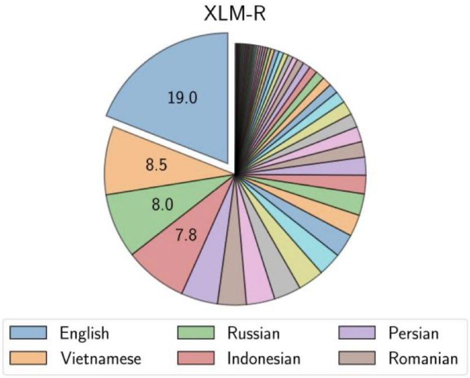


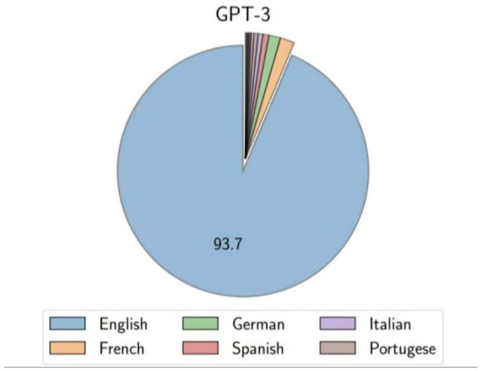


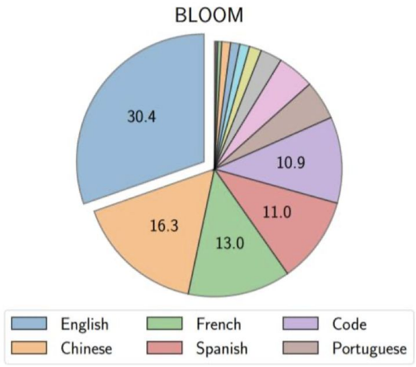
Multilingual LLMs: Data

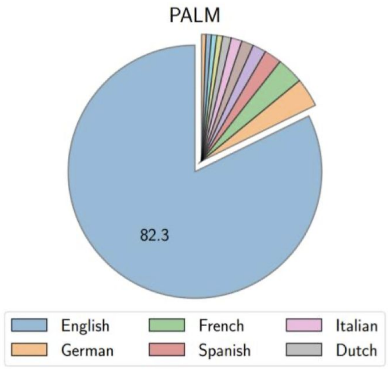
How to balance data?

Continue collecting!









Multilingual LLMs: Data

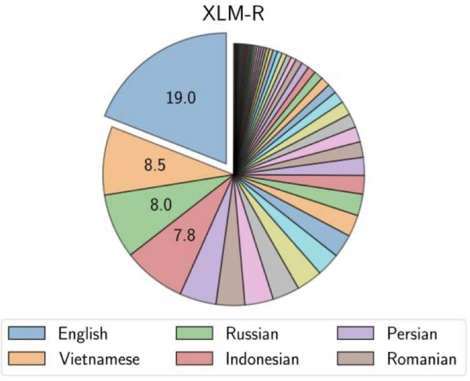
How to balance data?

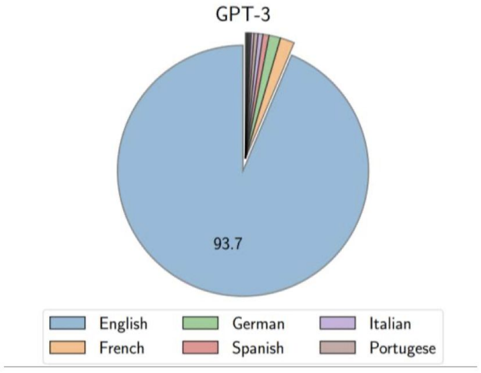
Continue collecting!

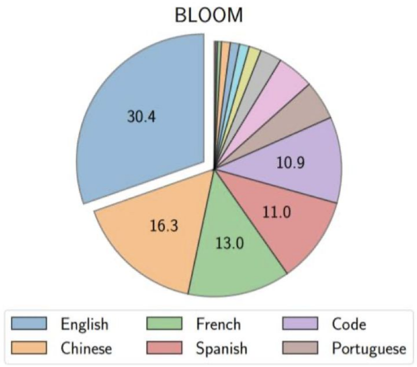
Popular approach is up-sampling of low-resource languages.

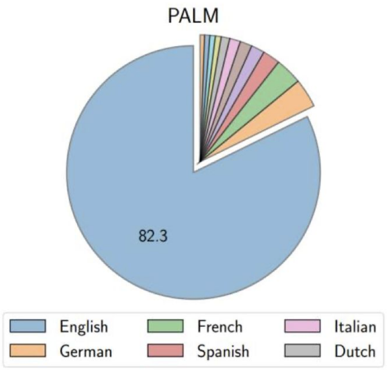
$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}$$

UNIMAX: Even stronger up-sampling To match underrepresented langs.



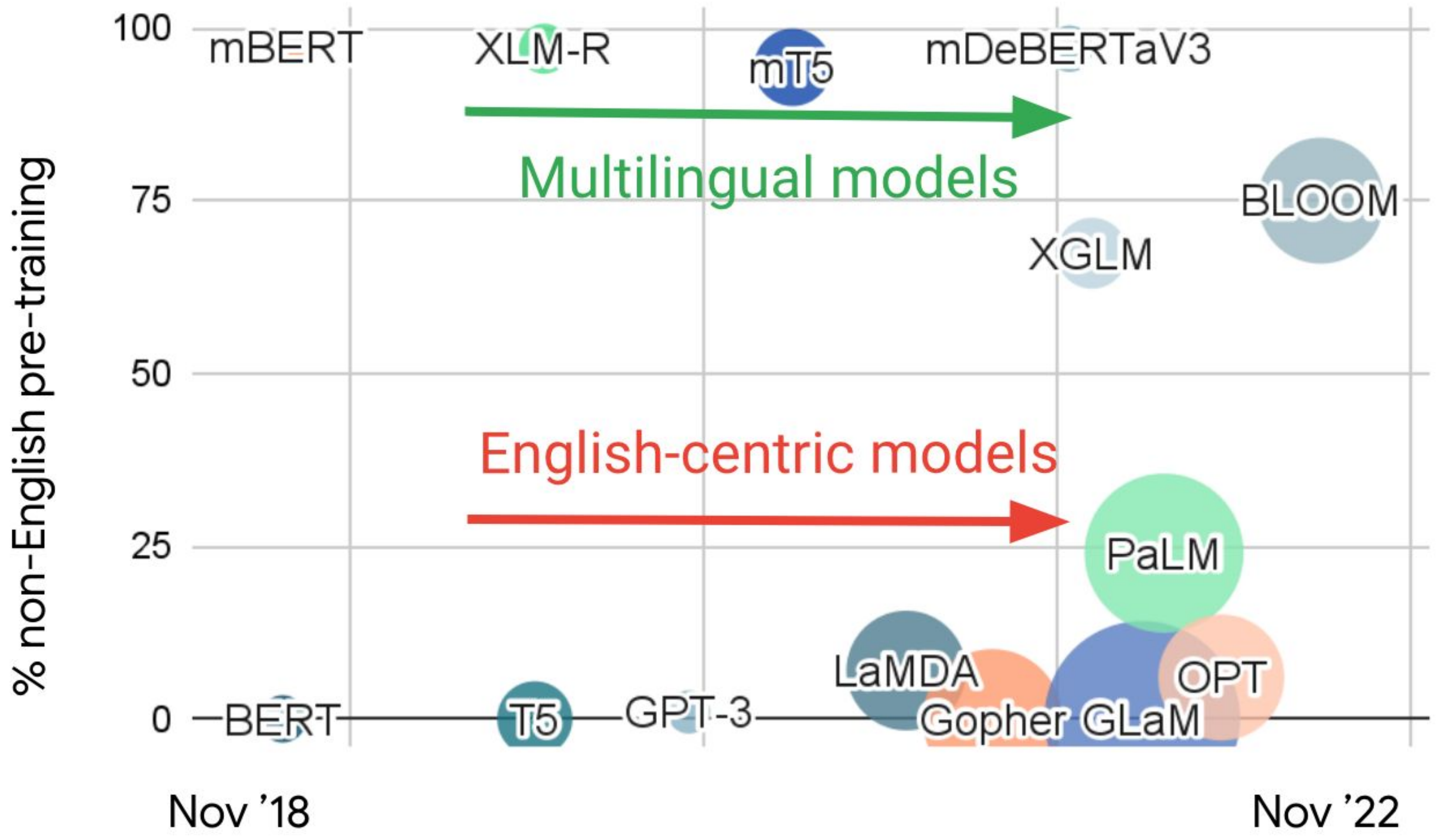






Multilingual LLMs: Data

source: www.ruder.io



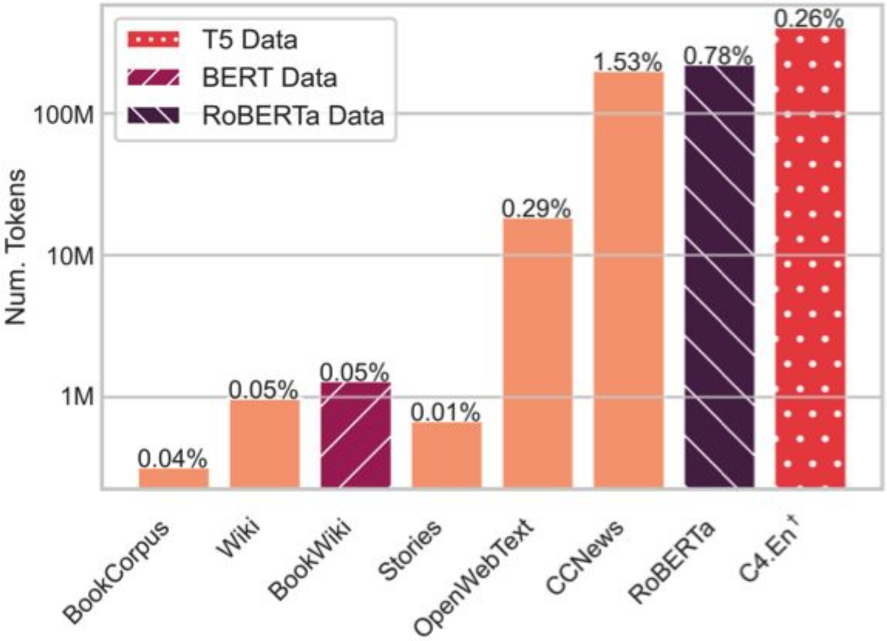
Do Monolingual Models Exist?

source: https://arxiv.org/pdf/2204.08110

Monolingual models show some relatively good multilingual capabilities.

One explanation of this phenomena are contaminations:

i.e. every monolingual model always contain a some samples from other languages.

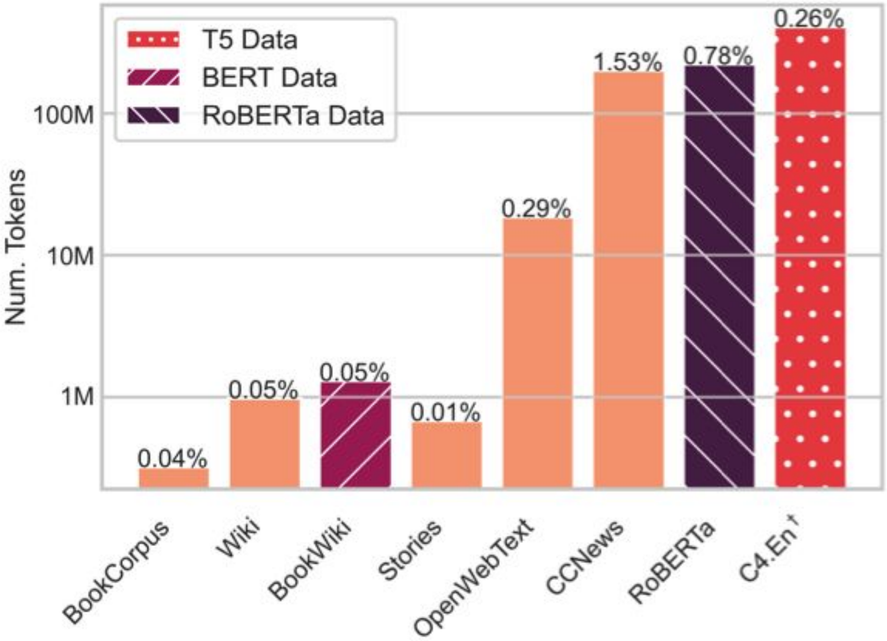


Share of non-English data in English corpora

Do Monolingual Models Exist?

source: https://arxiv.org/pdf/2204.08110

Monolingual models show some relatively good multilingual capabilities.



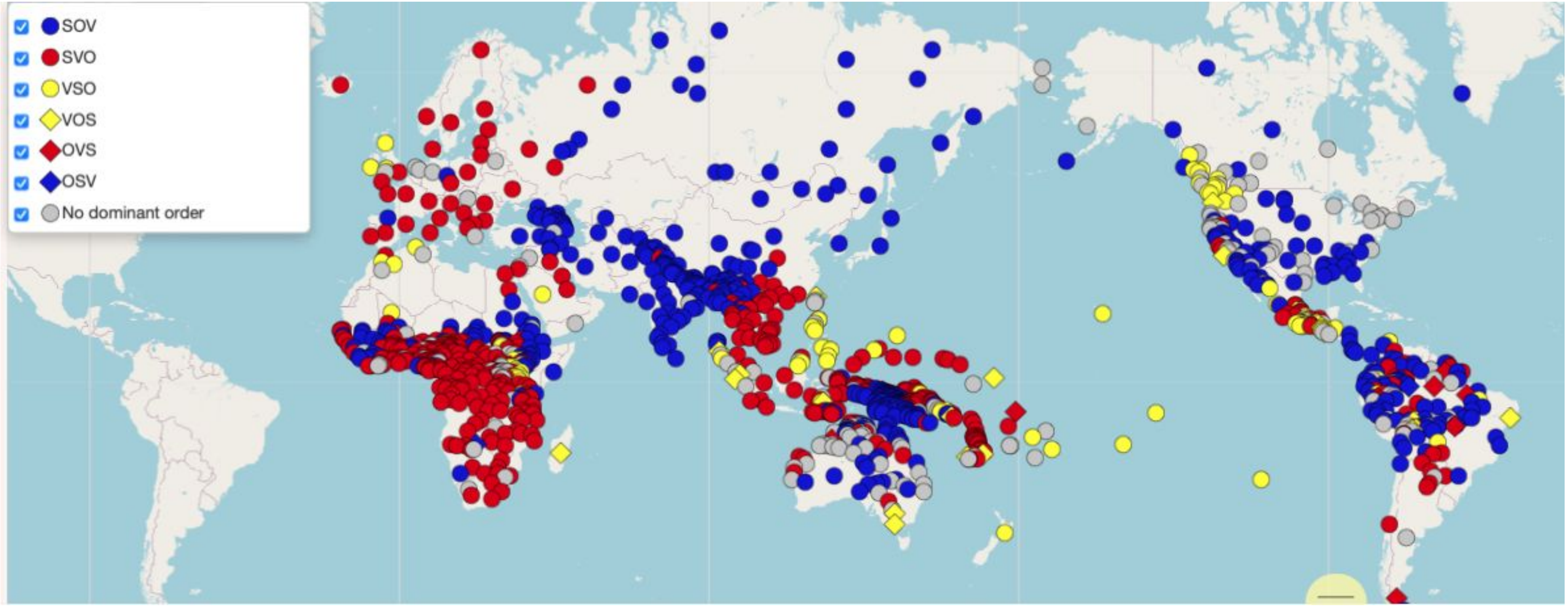
Share of non-English data in English corpora

Languages Are Different

Multilingual LLMs: Variability of Languages

source: https://wals.info

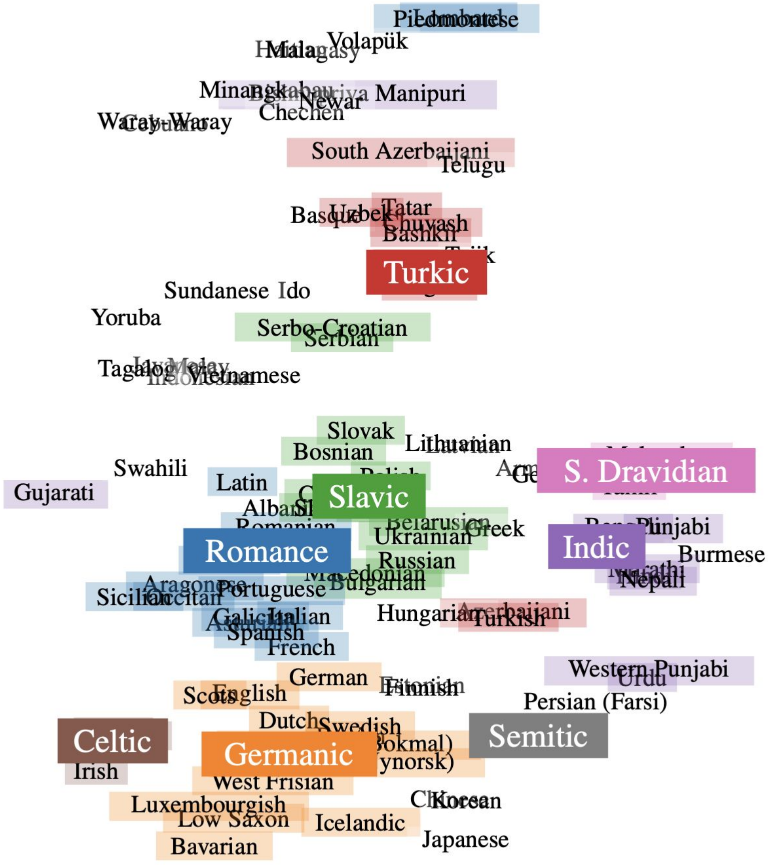
Order of Subject, Object, Verb throughout languages



Multilingual LLMs: Variability of Languages

source: https://arxiv.org/pdf/1911.03310

Transfer tends to be better between pairs of typologically related languages

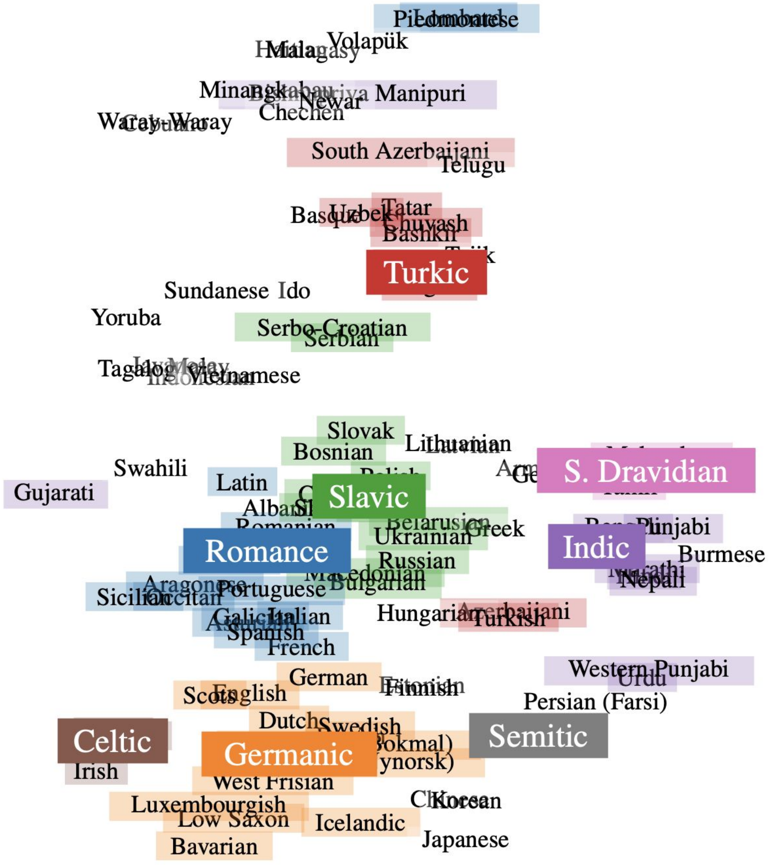


Multilingual LLMs: Variability of Languages

source: https://arxiv.org/pdf/1911.03310

Transfer tends to be better between pairs of typologically related languages

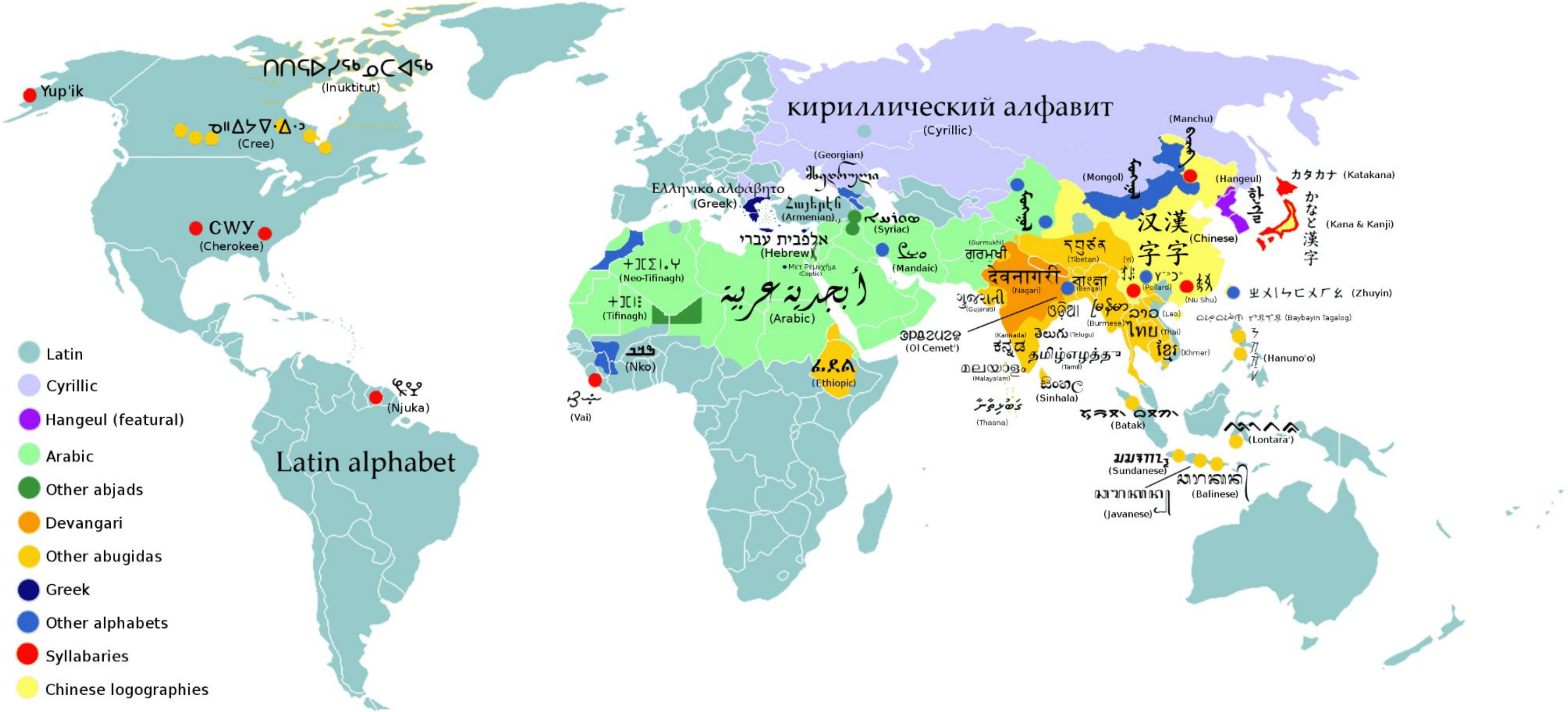
Another issue to consider: writing system



Remember about Tokenization

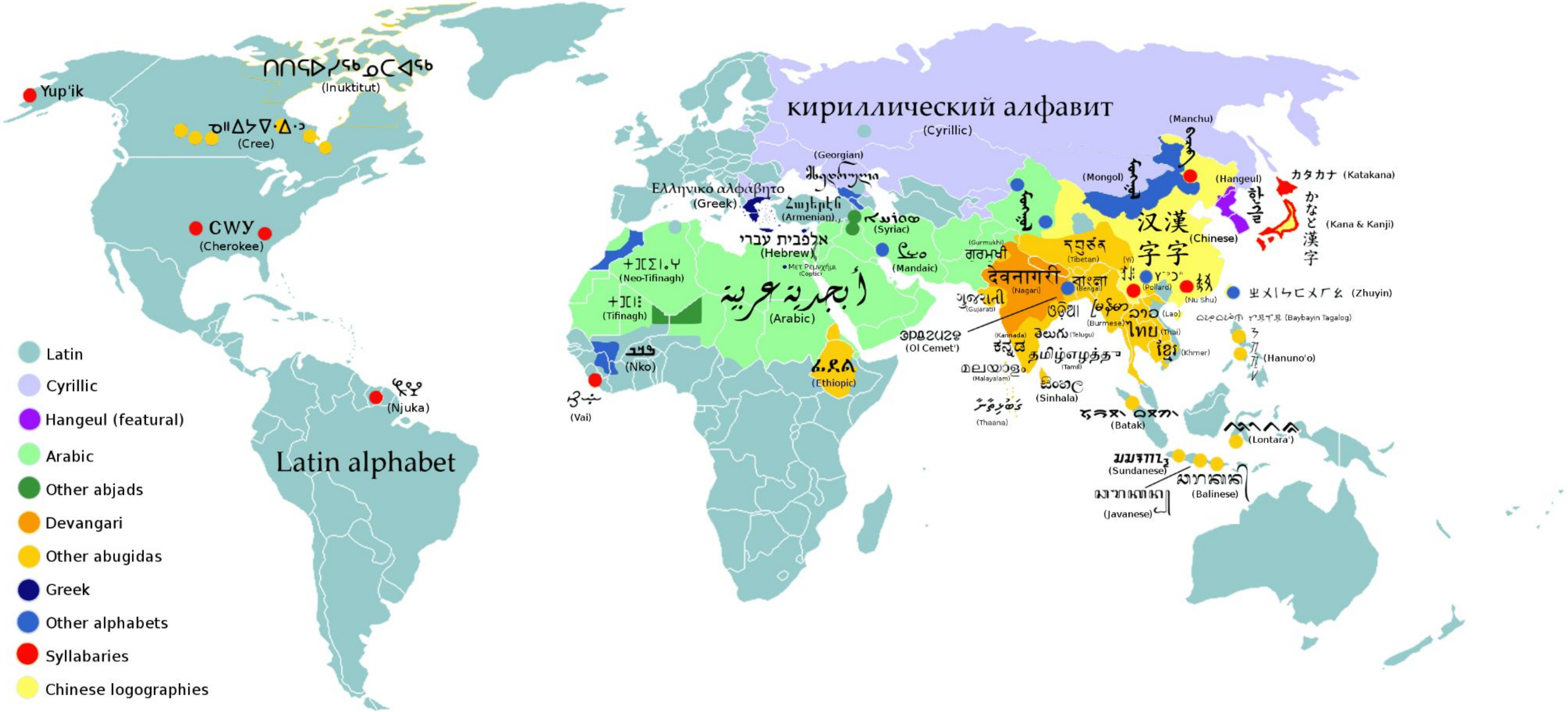
Multilingual LLMs: Tokenization

source: wikipedia.org



Multilingual LLMs: Tokenization

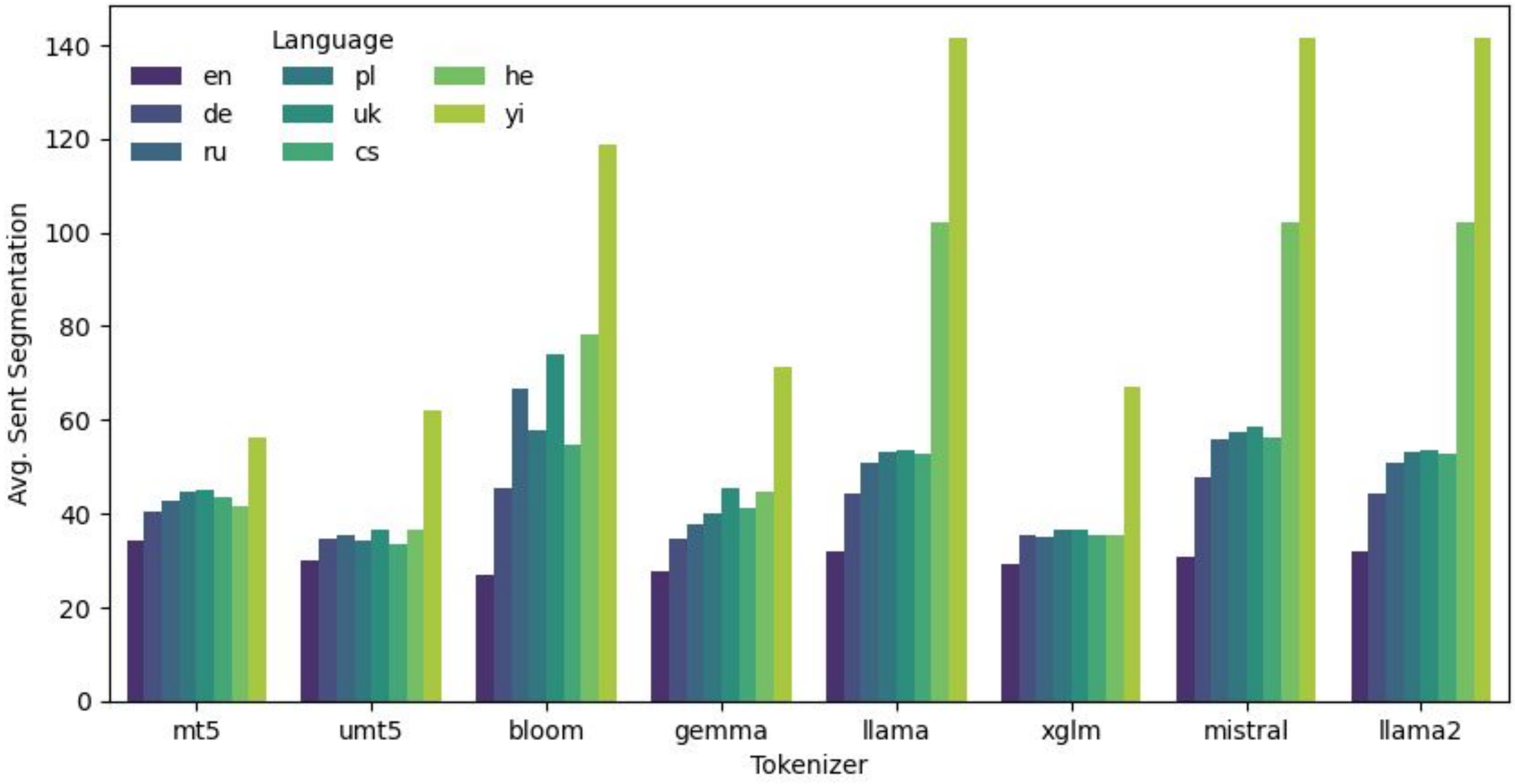
source: wikipedia.org



Beware of out-of-vocabulary tokens or [UNK]s 😨😨😨

Multilingual LLMs: Tokenization

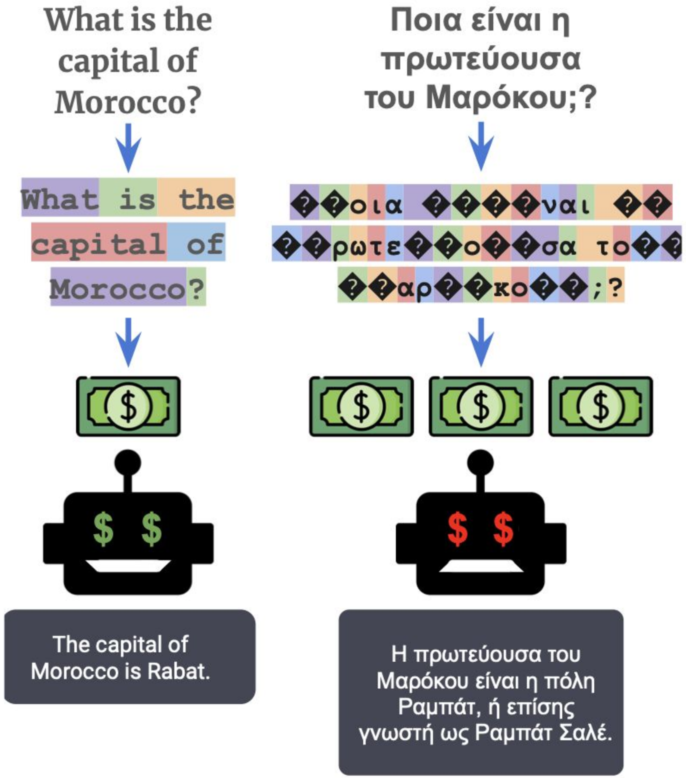
source: own analysis



Multilingual LLMs: Tokenization

source: https://arxiv.org/pdf/2305.13707

- **Oversegmentation** increases computation cost for low-resource languages
- Splitting into smaller chunks deteriorates performance
- Less fine-tuning examples fit the context window



What is the capital of Morocco?

Ποια είναι η πρωτεύουσα του Μαρόκου;?

The capital of Morocco is Rabat.

Η πρωτεύουσα του Μαρόκου είναι η πόλη Ραμπάτ, ή επίσης γνωστή ως Ραμπάτ Σαλέ.

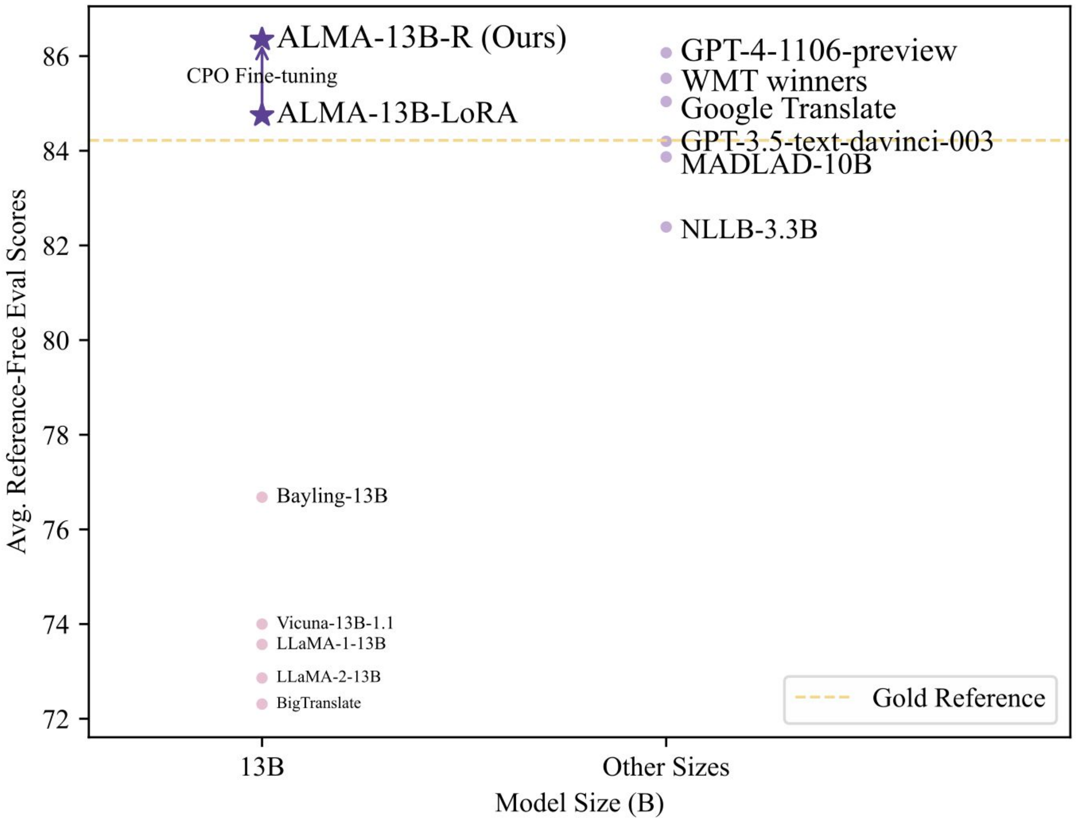
LLMs for Machine Translation

LLMs for Machine Translation

source: https://arxiv.org/pdf/2401.08417

Closed models (GPT-4) perform quite well

Open models need multilingual data and instruction tuning (ALMA, TowerInstuct)



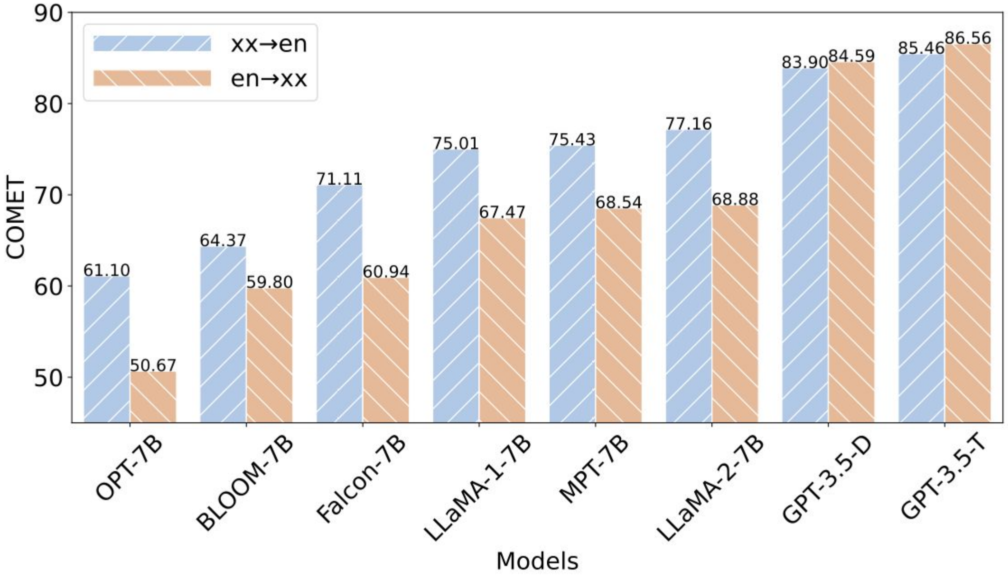
LLMs for Machine Translation: Evaluation

source: https://arxiv.org/pdf/2401.08417

Outperform dedicated MT models

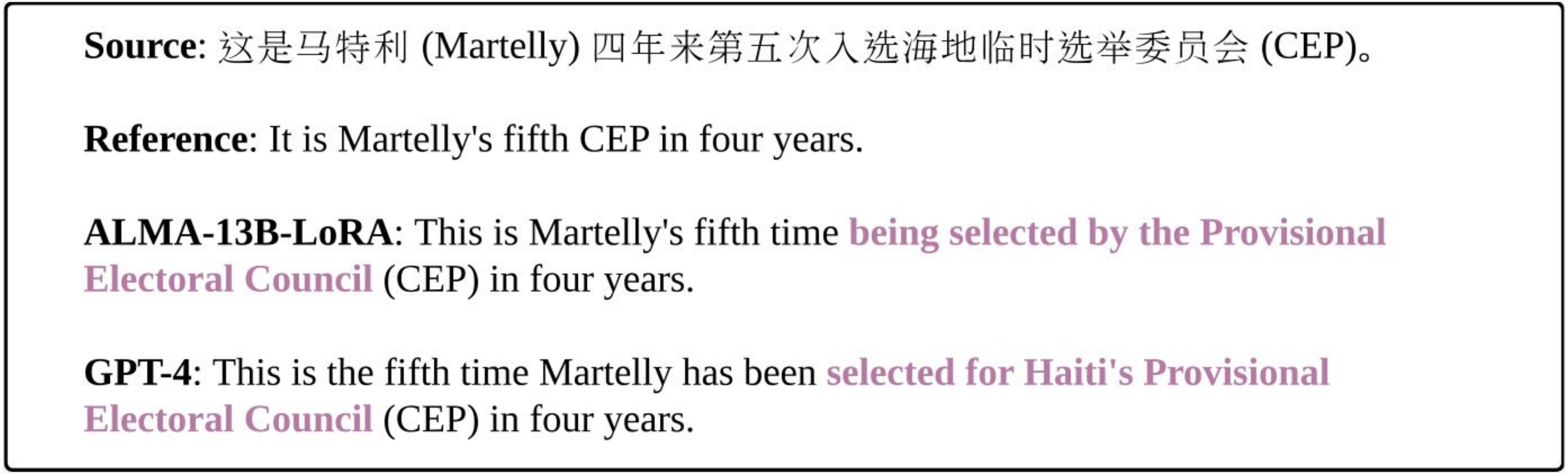
Better in translation to English than from English translation

Perform well on tasks other than machine translation



LLMs for Machine Translation: Evaluation

Can be better than reference from multilingual parallel data



Source: 这是马特利 (Martelly) 四年来第五次入选海地临时选举委员会 (CEP)。

Reference: It is Martelly's fifth CEP in four years.

ALMA-13B-LoRA: This is Martelly's fifth time **being selected by the Provisional Electoral Council** (CEP) in four years.

GPT-4: This is the fifth time Martelly has been **selected for Haiti's Provisional Electoral Council** (CEP) in four years.

Questions?