

# Multilingual NLP

*Author: Matthieu Futeral*

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

# Evaluating NMT - BLEU score

**English**

*I cannot make it today.*

**French**

*Je ne peux pas venir aujourd'hui.*

How to evaluate if French translation is a good one?



# Evaluating NMT - BLEU score

**English**

*I cannot make it today.*

**French**

*Je ne peux pas venir aujourd'hui.*

How to evaluate if French translation is a good one?  
**BLEU Score**



# Evaluating NMT - BLEU score

English

*I cannot make it today.*

French (possible translations)

*Je ne peux pas venir aujourd'hui.*

*Je ne peux pas me déplacer aujourd'hui.*

*Je ne peux pas être présent aujourd'hui.*

**BLEU Score = Bilingual Evaluation Understudy score**

***How does it work?***

## Evaluating NMT - BLEU score

BLEU Score evaluates outputs of a translation system comparing them to the references (regardless of the input).

- Given an output translation, first count the number of 1-grams in the output translation that are in the reference translations (threshold by the maximum number of times the 1-gram is in one ref.)
- Do the same for 2, 3 and 4 grams.
- Compute the geometric mean of BLEU-1 ... BLEU-4 to get the MEAN BLEU Score
- Multiply by a penalty length to get the final BLEU score.

# Evaluating NMT - BLEU score

## Example 1:

Ref 1: There is a cat on the mat.

Ref 2: The cat is on the mat.

Output: The cat is on the couch.

**BLEU-1:** The cat is on the couch . => 6/7

**BLEU-2:** The cat cat is is on on the the couch couch . => 4/6

**BLEU-3:** The cat is cat is on is on the on the couch the couch . => 3/6

**BLEU-4:** The cat is on cat is on the is on X the couch on the X couch . => 2/4

$$\text{BLEU} = \text{BP} * \exp(1 * \log(6/7) + \frac{1}{2} * \log(4/6) + \frac{1}{3} * \log(3/6) + \frac{1}{4} * \log(2/4)) \quad (\text{BP} = 1)$$

# Evaluating NMT - BLEU score

## Example 2:

Ref 1: There is a cat on the mat.

Ref 2: The cat is on the mat.

Output: Cat on mat.

**BLEU-1:** Cat    on    mat    .                          => 4/4

**BLEU-2:** Cat on    on mat    mat .                          => 3/3

**BLEU-3:** Cat on mat        on mat .                          => 2/2

**BLEU-4:** Cat on mat .                                  => 1/1

**BLEU** = BP \* 1                          Here, length(Output) = 4 < length(min(ref)) = 7   => BP = exp(1 - 7/4)

**BLEU** = 0.4723

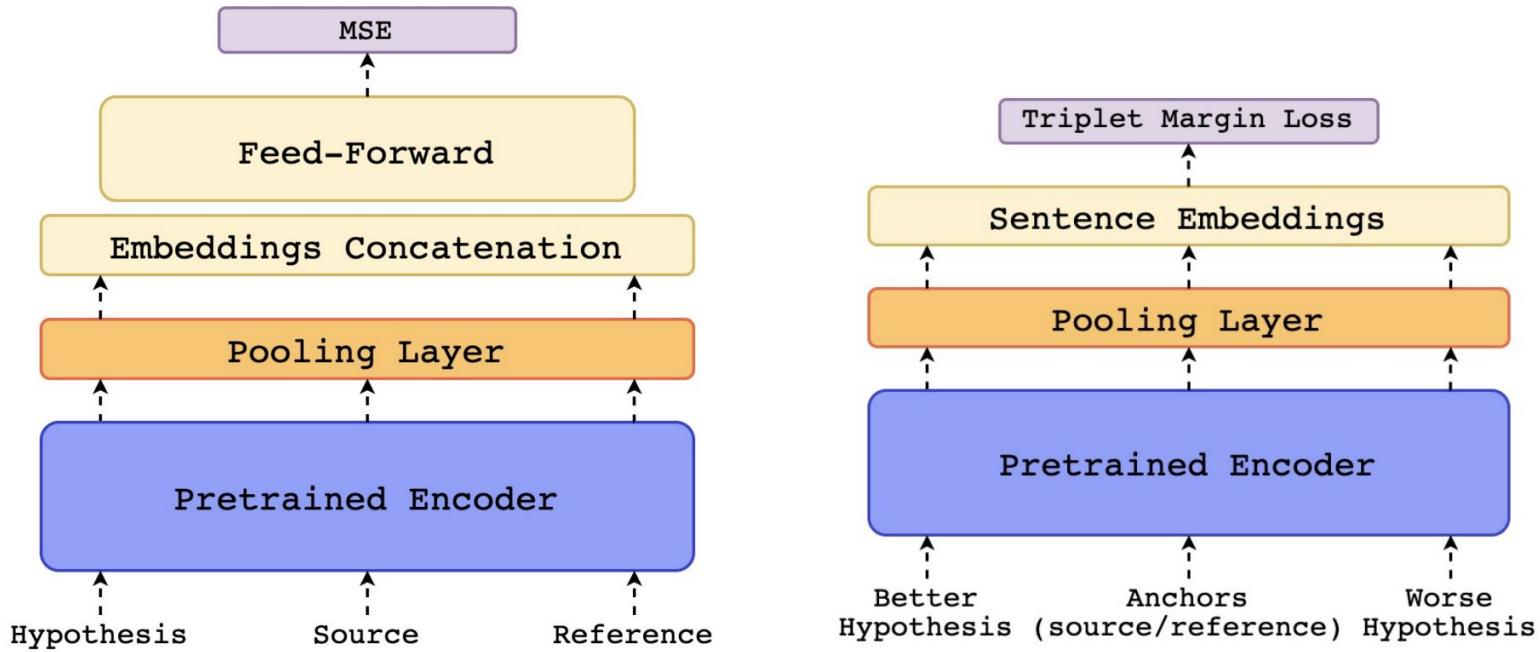
# Evaluating NMT - BLEU score

## Limitations:

- Higher score for long sentences than short (BP = 1)
- Just a matching word evaluation => a synonym not included in the references is considered an error in the same way as any other words.
- No notion of grammar.

Solution: **COMET**

# Evaluating NMT - COMET score, a neural based metric



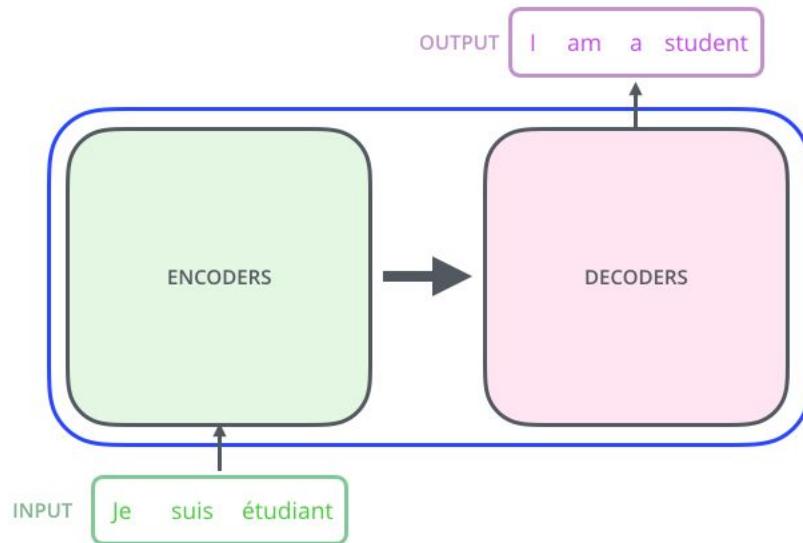
# Evaluating NMT - COMET score, a neural based metric

Metric	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
BLEU	0.364	0.248	0.395	0.463	0.363	0.333	0.469	0.235
CHRF	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241
YISI-1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355
BERTSCORE (default)	0.500	0.363	0.527	0.568	0.540	0.464	0.585	0.356
BERTSCORE (xlmr-base)	0.503	0.369	0.553	0.584	0.536	0.514	0.599	0.317
COMET-HTER	0.524	0.383	0.560	0.552	0.508	0.577	0.539	0.380
COMET-MQM	0.537	0.398	0.567	0.564	0.534	0.574	<b>0.615</b>	0.378
COMET-RANK	<b>0.603</b>	<b>0.427</b>	<b>0.664</b>	<b>0.611</b>	<b>0.693</b>	<b>0.665</b>	0.580	<b>0.449</b>

Comparing different NMT metric on the WMT 2019 Metrics Shared task.  
Score is Kendall's tau. Table from Rei et al. (2020)

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - **Sequence-to-Sequence models from scratch**
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

# Sequence-to-Sequence models - Bilingual



- Requires large amount of parallel data
- Needs one model for each language pair & direction

# Sequence-to-Sequence models - Multilingual

- Still requires large amount of parallel data
- Single model for all language directions
- Joint SentencePiece vocabulary

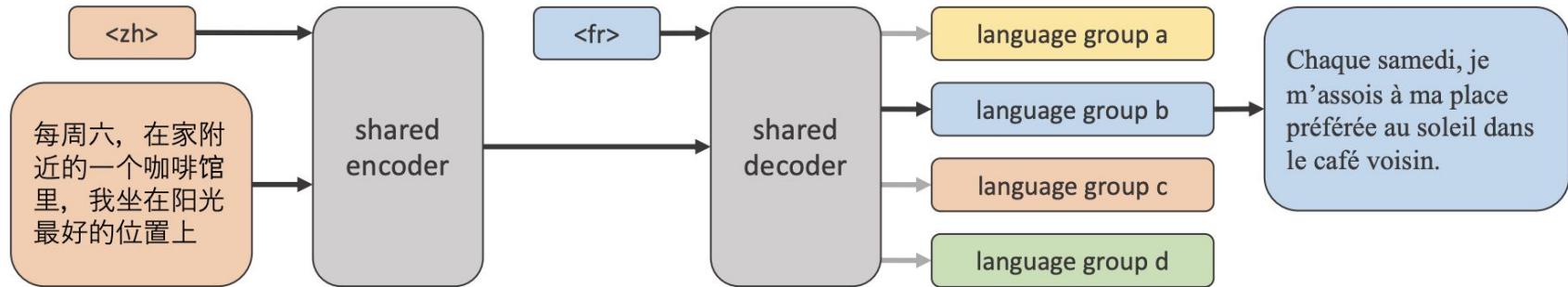


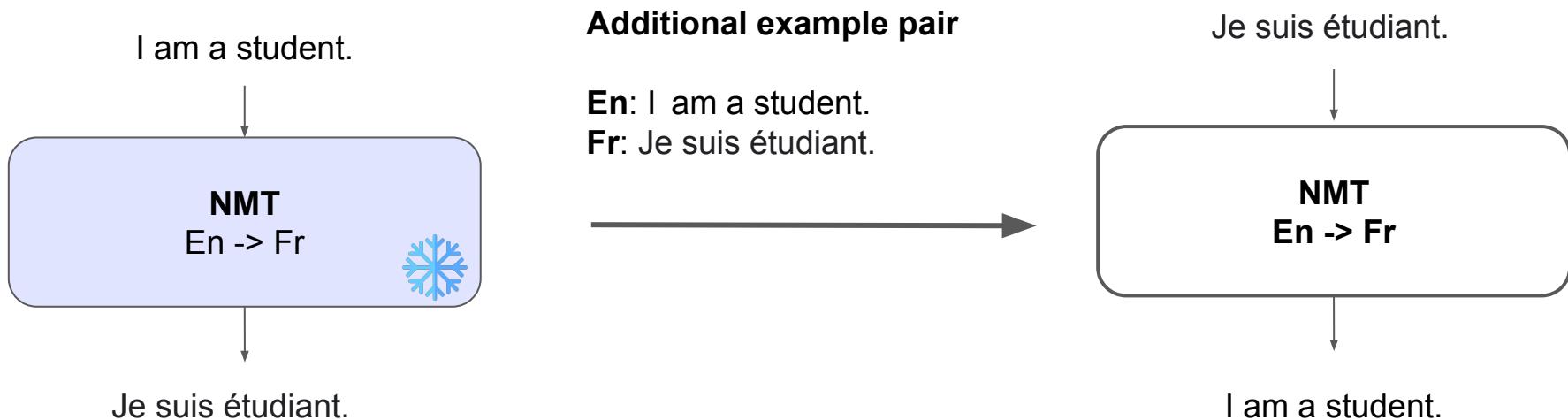
Illustration of the M2M-100 model from Fan et al. (2020)

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - **Leveraging monolingual data**
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

# Leveraging Monolingual data - Backtranslation

- Parallel data is scarce

What if we could use monolingual (cheaper to get)? **Backtranslation**



# Leveraging Monolingual data - Backtranslation

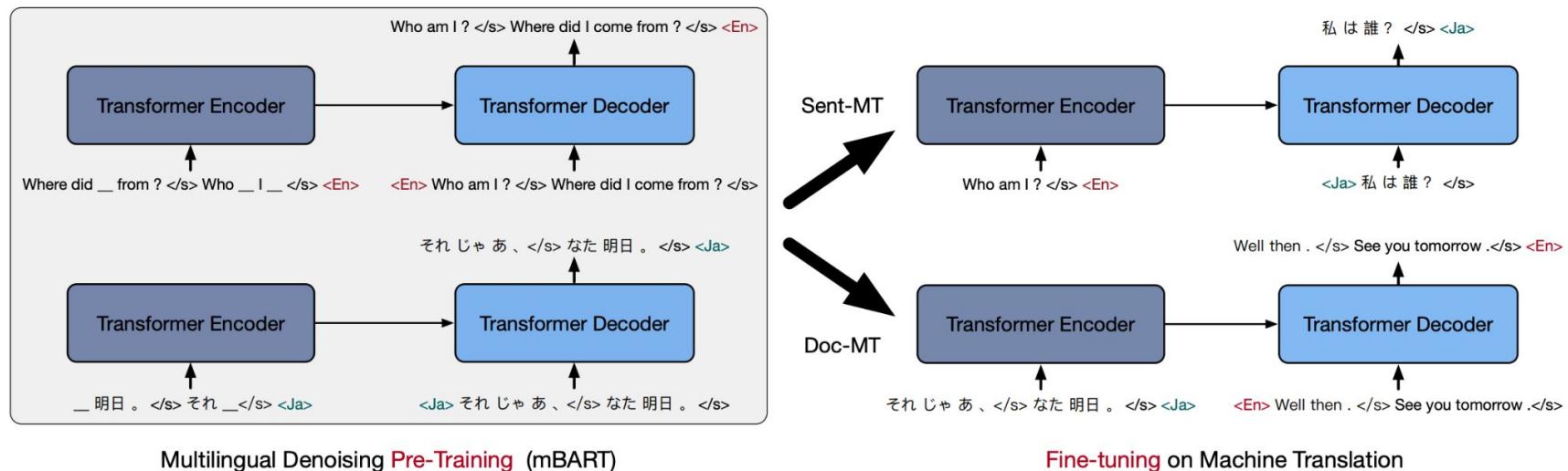
Backtranslation gives significant improvement!

<b>Model</b>	<b>BLEU score</b>
Transformer (baseline)	36.3
Transformer + Back-Translation	38.7

BLEU Results on IWSLT 2015 English -> Turkish from Cuong et al. (2020)

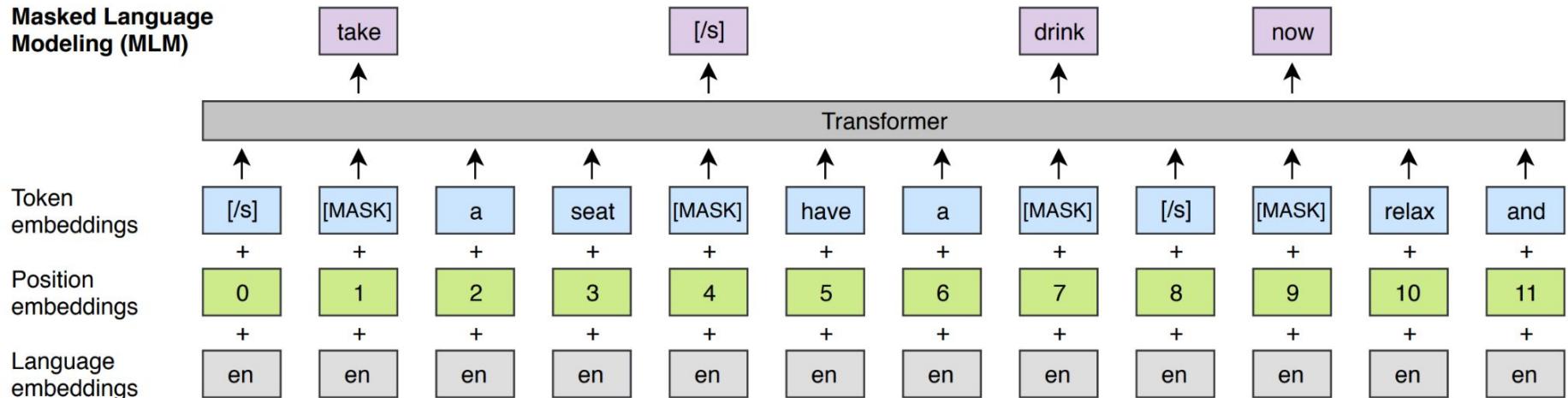
# Leveraging Monolingual data - Multilingual Pretraining mBART

- Pretraining on unlabelled data proved to be efficient in many tasks (e.g. BERT)



# Leveraging Monolingual data - Multilingual Pretraining XLM

- Another type of pretraining approach XLM



# Leveraging Monolingual data - Multilingual Pretraining

- Performances of pretraining

Model	Data	Fine-tuning		
		En→Ro	Ro→En	+BT
<b>Random</b>	None	34.3	34.0	36.8
<b>XLM (2019)</b>	En Ro	-	35.6	38.5
<b>MASS (2019)</b>	En Ro	-	-	39.1
<b>BART (2019)</b>	En	-	-	38.0
<b>XLM-R (2019)</b>	CC100	35.6	35.8	-
<b>BART-En</b>	En	36.0	35.8	37.4
<b>BART-Ro</b>	Ro	37.6	36.8	38.1
<b>mBART02</b>	En Ro	<b>38.5</b>	<b>38.5</b>	<b>39.9</b>
<b>mBART25</b>	CC25	37.7	37.8	38.8

BLEU score comparison of different pretraining approaches on  
En <-> Ro WMT 2016 test set from Liu et al. 2020

# Leveraging Monolingual data - Multilingual Pretraining

- Huge (resp. small) gains for low (resp. high) resource languages

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	WMT19		WMT19		IWSLT15		WMT17		IWSLT17		IWSLT17	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	<b>0.3</b>	<b>0.1</b>	<b>7.4</b>	<b>2.5</b>	<b>36.1</b>	<b>35.4</b>	<b>22.5</b>	<b>17.8</b>	<b>19.1</b>	<b>19.4</b>	<b>24.6</b>	<b>22.6</b>

Low/Medium resource MT - BLEU score comparison between mBART pretraining and random init from Liu et al. (2020)

Languages	Cs	Es	Zh	De	Ru	Fr
Size	11M	15M	25M	28M	29M	41M
Random	16.5	33.2	<b>35.0</b>	<b>30.9</b>	<b>31.5</b>	<b>41.4</b>
mBART25	<b>18.0</b>	<b>34.0</b>	33.3	30.5	31.3	41.0

High resource MT - BLEU score comparison between mBART pretraining and random init on WMT test sets from Liu et al. (2020)

# Leveraging Monolingual data - Multilingual Pretraining

- Generalization to **Unseen** languages during pretraining

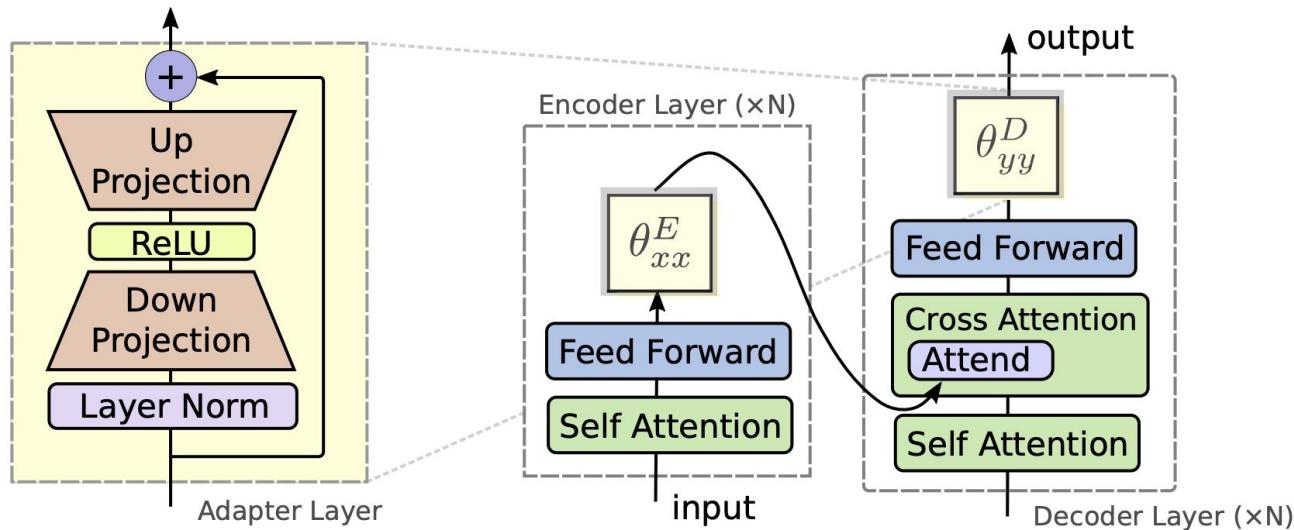
	<b>Monolingual</b>	<b>Nl-En</b>	<b>En-Nl</b>	<b>Ar-En</b>	<b>En-Ar</b>	<b>Nl-De</b>	<b>De-Nl</b>
<b>Random</b>	None	34.6 (-8.7)	29.3 (-5.5)	27.5 (-10.1)	16.9 (-4.7)	21.3 (-6.4)	20.9 (-5.2)
<b>mBART02</b>	En Ro	41.4 (-2.9)	34.5 (-0.3)	34.9 (-2.7)	21.2 (-0.4)	26.1 (-1.6)	25.4 (-0.7)
<b>mBART06</b>	En Ro Cs It Fr Es	43.1 (-0.2)	34.6 (-0.2)	37.3 (-0.3)	21.1 (-0.5)	26.4 (-1.3)	25.3 (-0.8)
<b>mBART25</b>	All	<b>43.3</b>	<b>34.8</b>	<b>37.6</b>	<b>21.6</b>	<b>27.7</b>	<b>26.1</b>

Generalization to unseen languages. BLEU scores comparison between mBART pretraining and random init. From Liu et al. (2020)

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - **Zero-shot Machine translation**
  - Learning a new language with LLMs
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

# Zero-shot Machine Translation - Adapters

Monolingual adapters to perform NMT on new language directions



Monolingual adapters are inserted in each layer.  $xx$  denotes a source language and  $yy$  a target language.

# “Zero-shot” Machine Translation - Performances of LLM

The case of CHATGPT (GPT4) from Jiao et al. (2023)

Translation Prompt	
TP1	Translate these sentences from [SRC] to [TGT]:
TP2	Answer with no quotes. What do these sentences mean in [TGT]?
TP3	Please provide the [TGT] translation for these sentences:

Translation prompts used

System	BLEU $\uparrow$	ChrF++ $\uparrow$	TER $\downarrow$
Google	31.66	57.09	56.21
DeepL	31.22	56.74	57.84
Tencent	29.69	56.24	57.16
ChatGPT w/ TP1	23.25	53.07	66.03
ChatGPT w/ TP2	24.54	53.05	63.79
ChatGPT w/ TP3	<b>24.73</b>	<b>53.71</b>	<b>62.84</b>

BLEU scores for the different prompts  
FLORES-101 En->Zh

# “Zero-shot” Machine Translation - Performances of LLM

The case of CHATGPT (GPT4) from Jiao et al. (2023)

Table 5: Performance of ChatGPT for translation robustness on domain-specific or noisy test data.

System	W19 Bio		W20 Rob2		W20 Rob3	
	De⇒En	En⇒Ja	Ja⇒En	De⇒En	De⇒En	
Google	37.83	29.72	19.21	42.91		
DeepL	37.13	26.25	19.83	41.29		
ChatGPT	33.22	22.36	18.34	44.59		

BLEU scores

# “Zero-shot” Machine Translation - Performances of LLM

The case of CHATGPT (GPT4) from Jiao et al. (2023)

System	De-En		Ro-En		Zh-En	
	⇒	⇐	⇒	⇐	⇒	⇐
Google	45.04	41.16	50.12	46.03	31.66	43.58
DeepL	49.23 <sub>(+9.3%)</sub>	41.46 <sub>(+0.7%)</sub>	50.61 <sub>(+0.9%)</sub>	48.39 <sub>(+5.1%)</sub>	31.22 <sub>(-1.3%)</sub>	44.31 <sub>(+1.6%)</sub>
Tencent	n/a	n/a	n/a	n/a	29.69 <sub>(-6.2%)</sub>	46.06 <sub>(+5.6%)</sub>
ChatGPT	43.71 <sub>(-2.9%)</sub>	38.87 <sub>(-5.5%)</sub>	44.95 <sub>(-10.3%)</sub>	24.85 <sub>(-46.0%)</sub>	24.73 <sub>(-21.8%)</sub>	38.27 <sub>(-12.1%)</sub>

System	De-Zh		Ro-Zh		De-Ro	
	⇒	⇐	⇒	⇐	⇒	⇐
Google	38.71	21.68	39.05	25.59	33.31	32.27
DeepL	40.46 <sub>(+4.5%)</sub>	22.82 <sub>(+5.2%)</sub>	38.95 <sub>(-0.2%)</sub>	25.39 <sub>(-0.7%)</sub>	35.19 <sub>(+5.6%)</sub>	34.27 <sub>(+6.1%)</sub>
Tencent	40.66 <sub>(+5.0%)</sub>	19.44 <sub>(-10.3%)</sub>	n/a	n/a	n/a	n/a
ChatGPT	34.46 <sub>(-10.9%)</sub>	19.80 <sub>(-8.6%)</sub>	30.84 <sub>(-21.0%)</sub>	19.17 <sub>(-25.0%)</sub>	33.38 <sub>(+0.2%)</sub>	29.89 <sub>(-7.3%)</sub>

BLEU scores on FLORES-101

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - **Learning a new language with LLMs**
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

# Learning a new language with LLMs

Are LLMs able to learn a new language based on a grammar book?

*Tanzer et al. (2024) => A Benchmark for Learning to Translate a new Language from one Grammar Book*

Research questions:

- Do LLMs really learn to perform new tasks by adaptation, or does adaptation simply draw out capabilities that the model had already learned?
- Does scaling pretraining data just improve performance because it implicitly scales up in-domain training data for every task?

Use case: Translation between English and Kalamang (spoken by 200 people on a small island in Indonesian Papua)

# Learning a new language with LLMs

Are LLMs able to learn a new language based on a grammar book?

*Tanzer et al. (2024) => A Benchmark for Learning to Translate a new Language from one Grammar Book*

Research questions:

- Do LLMs really learn to perform new tasks by adaptation, or does adaptation simply draw out capabilities that the model had already learned?
- Does scaling pretraining data just improve performance because it implicitly scales up in-domain training data for every task?

Use case: Translation between English and Kalamang (spoken by 200 people on a small island in Indonesian Papua)

# Learning a new language with LLMs

## Setup:

*Learning from explanation.* Grammar book to translate from English to Kalamang (new types of learning in comparison to feeding large quantities of parallel text data into the model).

- a) Grammar book
- b) List of bilingual words
- c) Short list of Kamalang-English translations

No context (-)

Wordlist Context (W)

Paired Sentence Context (S)

Grammar Book Context (G)

Neural Machine Translation

### 13.1.5 With give-constructions

Give-constructions (§12.2.1.2) are made with a zero morpheme ‘give’. They may and frequently do occur without any other verb in the clause. However, they also occur in complex predicates with predicate linker =i. The verb marked with =i precedes the recipient. The zero morpheme ‘give’ comes after the recipient, which makes these discontinuous complex predicates. The verbs only share their subject, and the recipient comes between the two verbs. The theme (pandanus leaf in the first example and fish in the second) is the direct object of both verbs.

- (31) *naman=a padanual=at rep=i ka*  $\emptyset$   
who=FOC pandanus=OBJ get=PLNK 2SG give  
'Who got pandanus [leaf] and gave it to you?'
- (32) *an toni kuru ma yap=i sontum=ki*  $\emptyset$   
1SG say bring move\_landwards divide=PLNK person=BEN give  
'I said bring it here and divide it among people.'

Figure 1: An excerpt from *A grammar of Kalamang* (Visser, 2022). Examples with interlinear gloss are interspersed among English explanations.

# Learning a new language with LLMs

Grammar Book Context (G): how to feed the grammar book context into the LLM?

(i) Retrieve passages from the book that are similar to the input sentence (512 tokens)

- Either similarity in the embedding space  $G^e$
- Or similarity with longest substring matching  $G^s$

(ii) Manually curate subset of the books and feed the whole context into the LLM

- Either medium ( $\sim 50K$  tokens) =>  $G^m$
- Or large ( $\sim 100K$  tokens) =>  $G^l$

# Learning a new language with LLMs

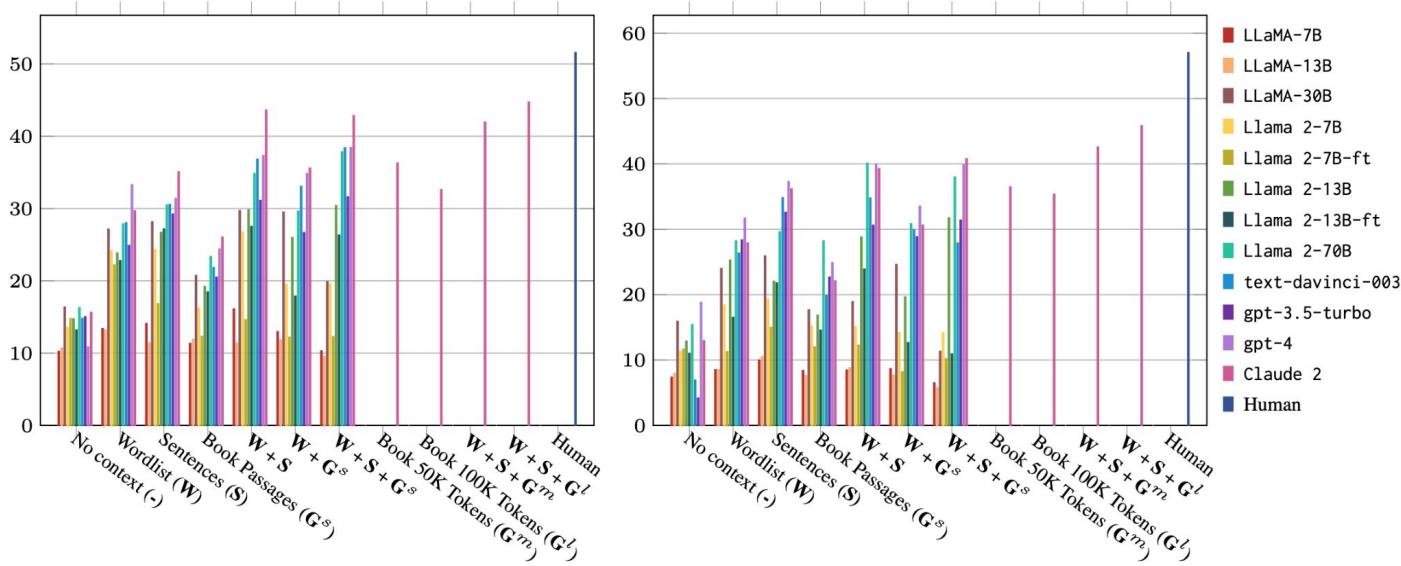


Figure 3: **chrF scores for kgv→eng (left) and eng→kgv (right) translation across experimental settings.** See Section 4.1.1 for details on the models in the legend; -ft represents finetuning on the grammar book text. See Section 4.1.2 for details on the provided context;  $W$  represents word list entries,  $S$  sentence pairs,  $G^s$  grammar book excerpts,  $G^m \sim 50K$  grammar book tokens,  $G^l \sim 100K$  grammar book tokens, and + combinations thereof. We see that quality depends both on the underlying model and the provided reference materials, with the best results coming from Claude 2 in the  $W + S + G^l$  setting. Human performance considerably exceeds all model baselines.

# Learning a new language with LLMs

Key takeaways:

- **Effect of scale:** Models that are larger and trained on more data tend to perform better, e.g., across the LLaMA and Llama 2 families.
- **Effect of context:** Biggest boost is observed when introducing parallel sentences (S) + bilingual word pairs (W). Grammar books helps improve a bit the results when combined with S + W. (See Claude results)
- **Comparison to human baseline:** The human baseline considerably outperforms all model baselines, both in quantitative (+6.9 chrF and +11.2 chrF) and qualitative terms.

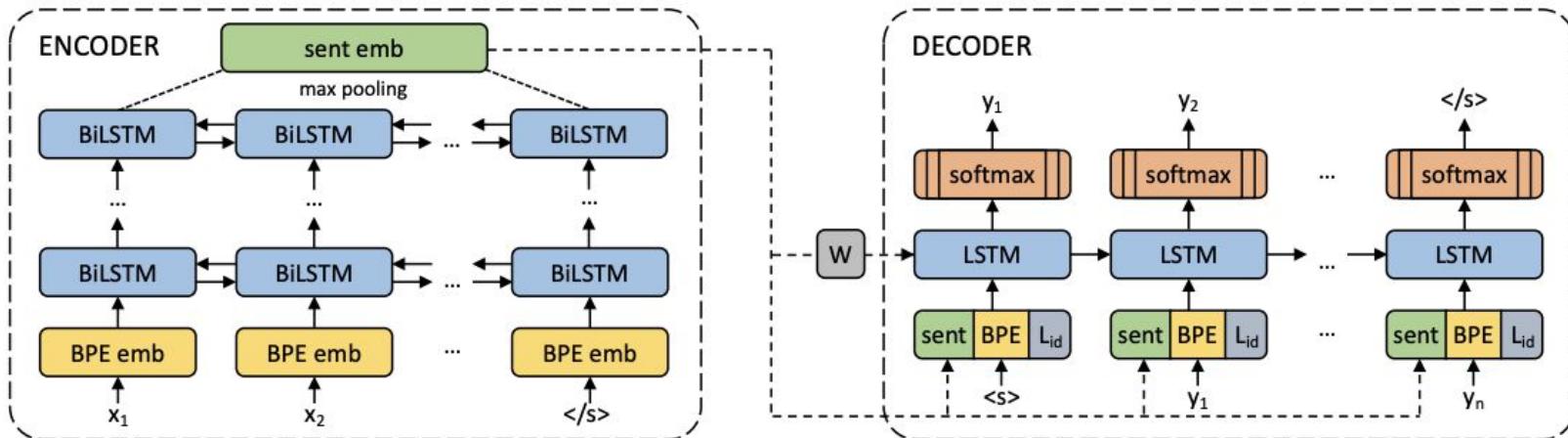
- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - **Data-centric approach**
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

## Data-centric approach

- Previous methods led to improvement in medium/low resource settings but performances are still limited by the lack of data.
- Could we mine translation pairs for any language pairs to increase the amount of parallel data available?

# Data-centric approach - LASER

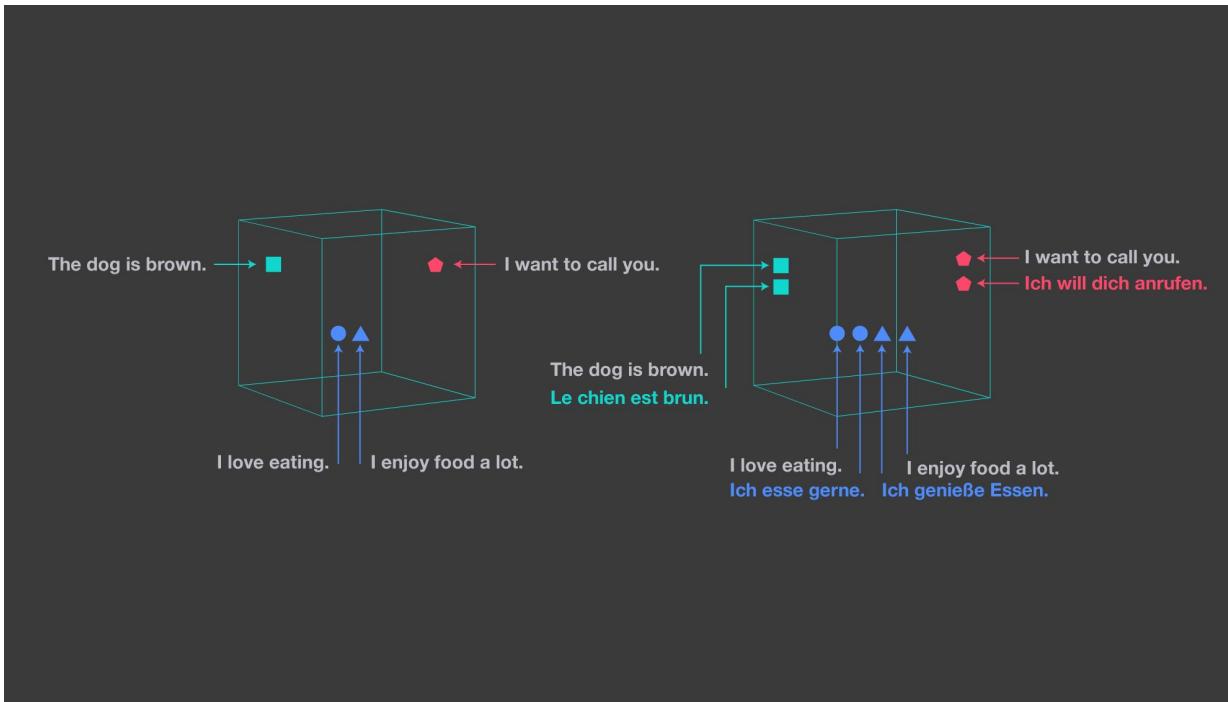
## Language Agnostic SEntence Representation - LASER



LASER modelling. The sentence embedding is what matters here.

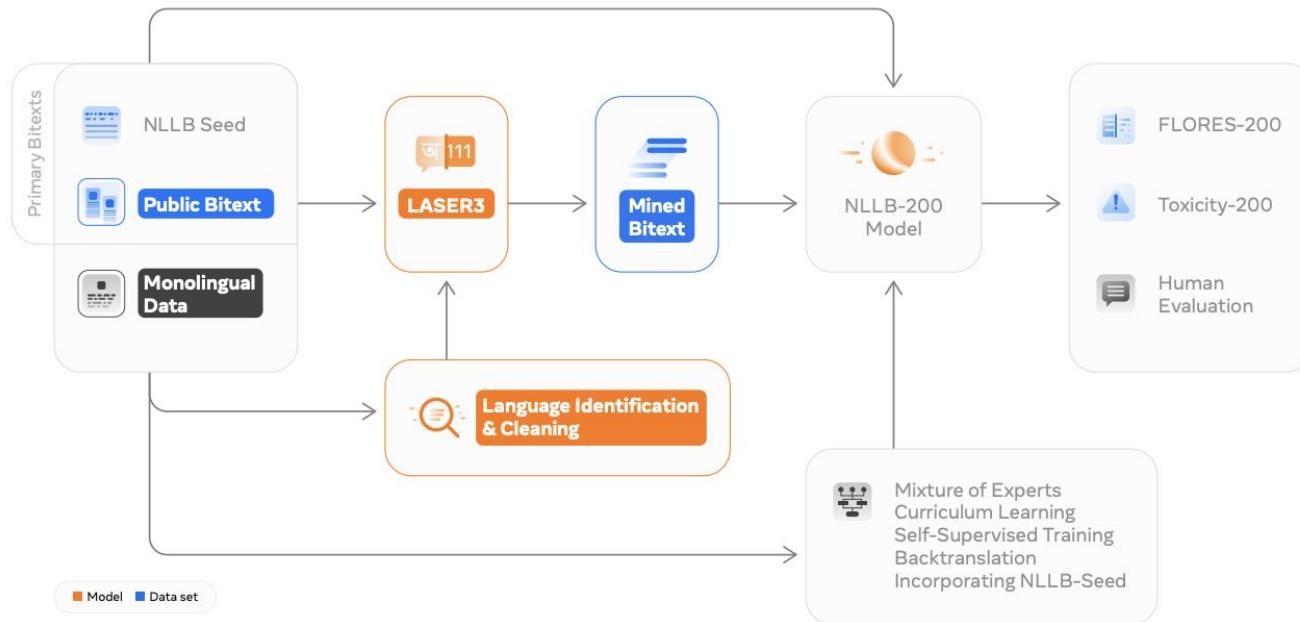
# Data-centric approach - LASER

Language Agnostic SEntence Representation - LASER



# Data-centric approach - NLLB

## No Language Left Behind (NLLB) approach



NLLB approach - Using LASER to mine bitext data from CommonCrawl

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - Data-centric approach
  - **Document-level Machine translation**
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

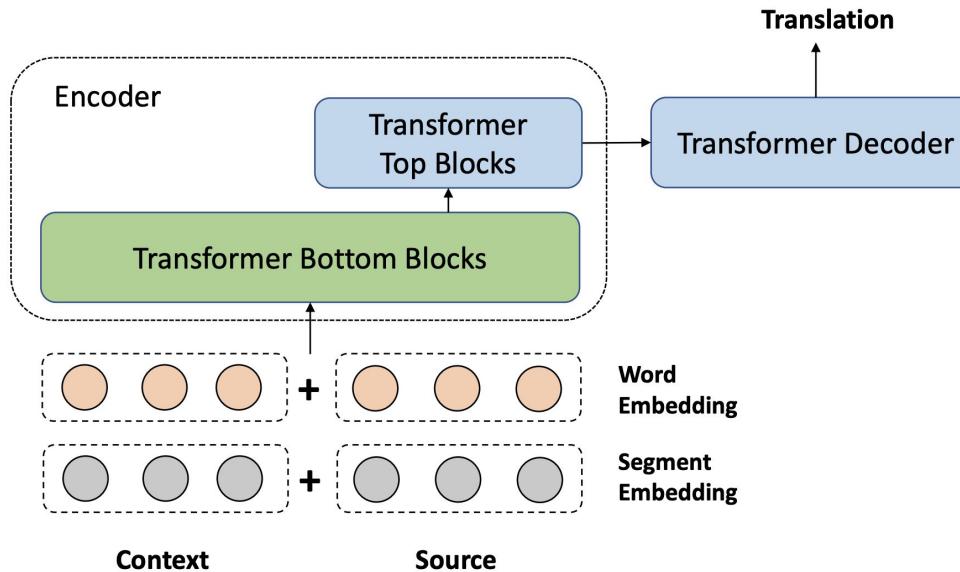
## Document-level Machine Translation

- Most works on NMT focuses on the translation of (short) sentences.
- Many tasks require to translate a whole document.
- 2 simple baselines:
  - **Translate sentences independently =>** Low document coherence
  - **Concatenate sentences =>** High computational cost + performance downgrades with long context

# Document-level Machine Translation - Concatenation methods

Different ways to concatenate contexts:

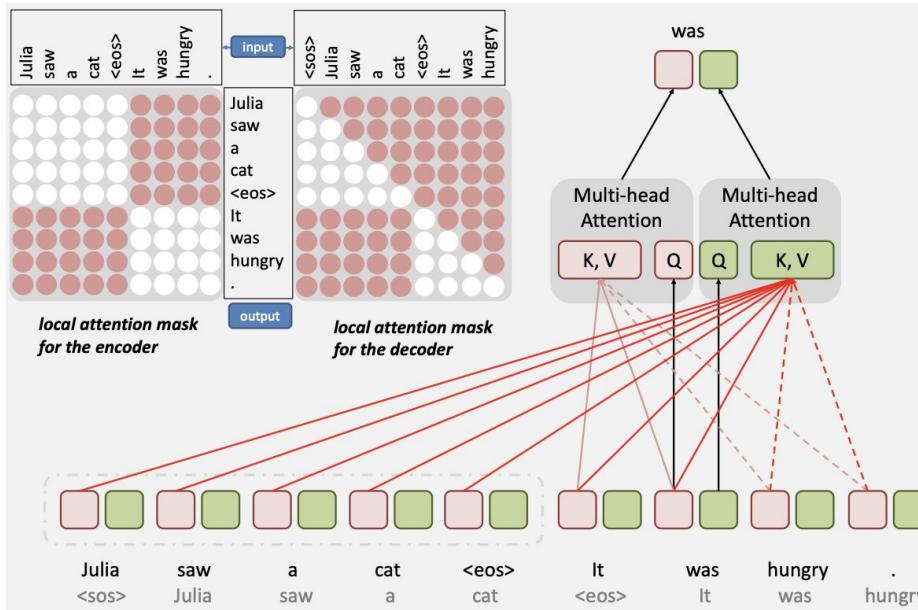
- Concatenation + self-attention:



# Document-level Machine Translation - Concatenation methods

Different ways to concatenate contexts:

- Long-Short term attention:



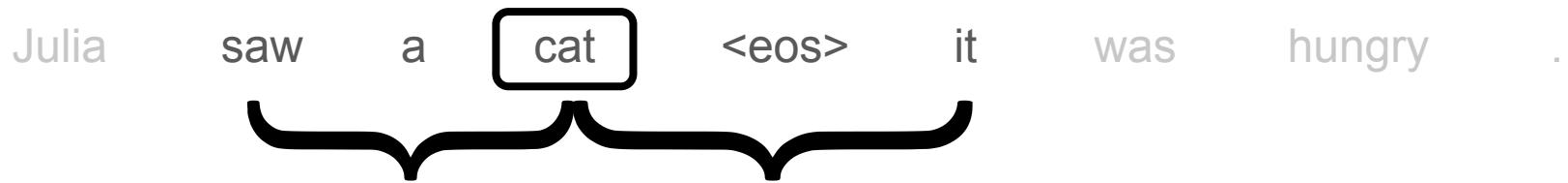
# Document-level Machine Translation - Modifying Attention

Herold et al. (2023) proposes window attention.

Julia saw a cat <eos> it was hungry .

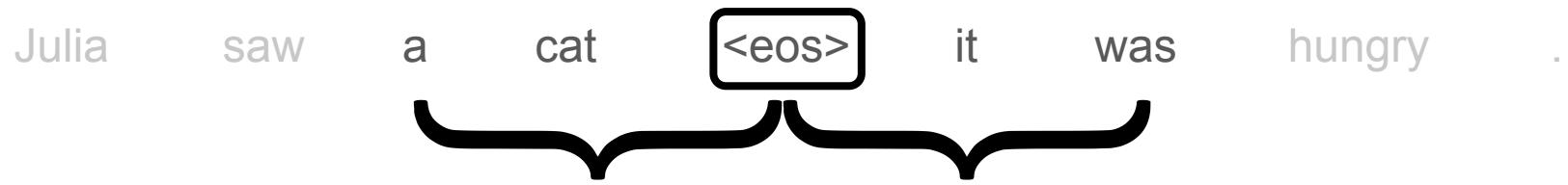
# Document-level Machine Translation - Modifying Attention

Herold et al. (2023) proposes window attention.



# Document-level Machine Translation - Modifying Attention

Herold et al. (2023) proposes window attention.



# Document-level Machine Translation - Results

Model	Context	NEWS		TED		OS	
		newstest2018		tst2017		test	
		BLEU	TER	BLEU	TER	BLEU	TER
sent.-level (external)	0 sent.	†32.3	-	‡33.4	-	*37.3	-
sent.-level (ours)		32.8	49.0	34.2	46.3	37.1	<b>43.8</b>
concat adj.	2 sent.	<b>33.4</b>	48.6	34.3	46.3	38.2	43.9
	1000 tok.	29.5	53.7	32.1	48.4	38.1	46.0
LST-attn	1000 tok.	30.0	53.1	29.8	54.5	<b>38.5</b>	45.1
window-attn	1000 tok.	33.1	<b>48.1</b>	<b>34.6</b>	<b>45.8</b>	38.3	44.4

- Neural Machine Translation
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

- Neural Machine Translation
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - Multicultural NLP

# Multilingual vs Monolingual models

Current NLP systems are English-centric (ChatGPT, BARD, Gemini etc.)

- What about the thousands spoken languages left?

Before diving into multilingual models, let's consider 2 baselines:

- 1) Translate-Train
- 2) Translate-Test

# Multilingual vs Monolingual models - Translate Train & Translate Test

Starting from a pretrained monolingual or multilingual LM

- **Zero-shot:**

Multilingual LM finetuned on English data and evaluated on other languages.

- **Translate Train:**

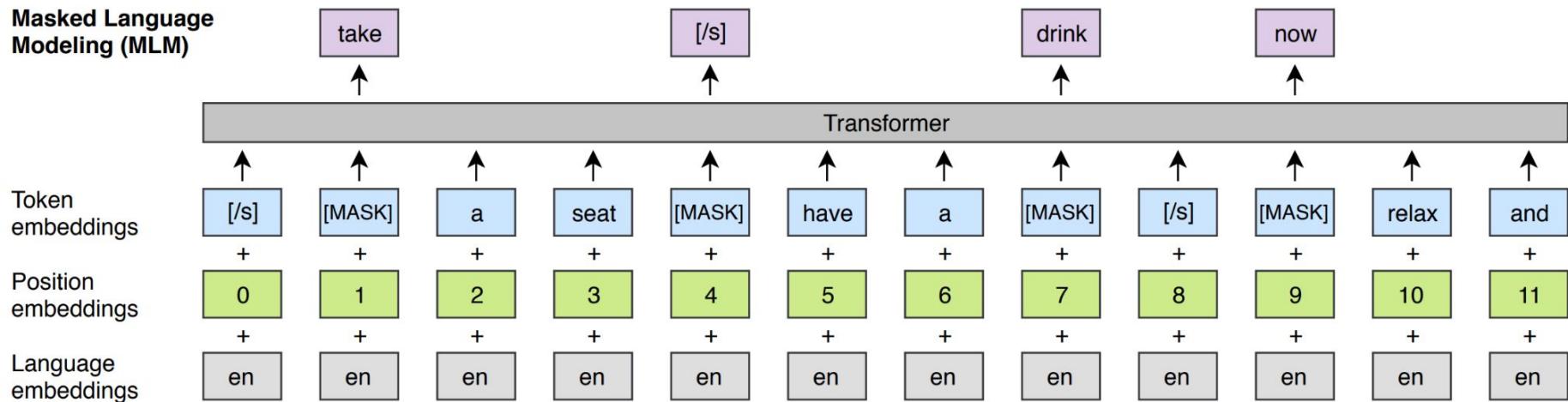
English training data is translated into target languages and the Multilingual LM is finetuned on the translated data

- **Translate Test:**

Translate the test sets into English and evaluate the multilingual LM or monolingual LM on the translated test sets.

# Multilingual vs Monolingual models - XLM-R

Multilingual MLM objective - trained on 100 languages



# Multilingual vs Monolingual models - XLM-R

Data is key!

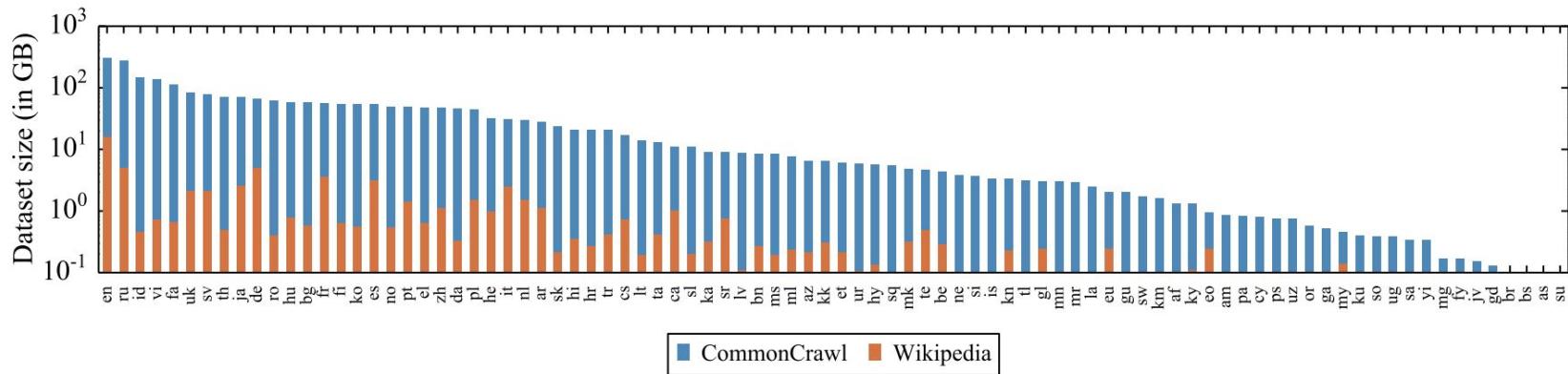


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

# Multilingual vs Monolingual models - XLM-R results

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
<i>Fine-tune multilingual model on English training set (Cross-lingual Transfer)</i>																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R <sub>Base</sub>	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	<b>89.1</b>	<b>84.1</b>	<b>85.1</b>	<b>83.9</b>	<b>82.9</b>	<b>84.0</b>	<b>81.2</b>	<b>79.6</b>	<b>79.8</b>	<b>80.8</b>	<b>78.1</b>	<b>80.2</b>	<b>76.9</b>	<b>73.9</b>	<b>73.8</b>	<b>80.9</b>
<i>Translate everything to English and use English-only model (TRANSLATE-TEST)</i>																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	<b>91.3</b>	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
<i>Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)</i>																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
<i>Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)</i>																			
Lample and Conneau (2019) <sup>†</sup>	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R <sub>Base</sub>	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	<b>89.1</b>	<b>85.1</b>	<b>86.6</b>	<b>85.7</b>	<b>85.3</b>	<b>85.9</b>	<b>83.5</b>	<b>83.2</b>	<b>83.1</b>	<b>83.7</b>	<b>81.5</b>	<b>83.7</b>	<b>81.6</b>	<b>78.0</b>	<b>78.1</b>	<b>83.6</b>

Table 1: **Results on cross-lingual classification.** We report the accuracy on each of the 15 XNLI languages and the average accuracy. We specify the dataset D used for pretraining, the number of models #M the approach requires and the number of languages #lg the model handles. Our *XLM-R* results are averaged over five different seeds. We show that using the translate-train-all approach which leverages training sets from multiple languages, *XLM-R* obtains a new state of the art on XNLI of 83.6% average accuracy. Results with <sup>†</sup> are from Huang et al. (2019).

# Multilingual vs Monolingual models in the target language

## The French case

CamemBERT, a Tasty French language model trained on OSCAR (Martin et al. 2020)

Model	F1
SEM (CRF) (Dupont, 2017)	85.02
LSTM-CRF (Dupont, 2017)	85.57
mBERT (fine-tuned)	87.35
CamemBERT (fine-tuned)	<u>89.08</u>
LSTM+CRF+CamemBERT (embeddings)	<b>89.55</b>

Table 3: **NER** scores on the FTB (best model selected on validation out of 4). Best scores in bold, second best underlined.

Model	Acc.	#Params
mBERT (Devlin et al., 2019)	76.9	175M
XLM <sub>MLM-TLM</sub> (Lample and Conneau, 2019)	<u>80.2</u>	250M
XLM-R <sub>BASE</sub> (Conneau et al., 2019)	80.1	270M
CamemBERT (fine-tuned)	<b>82.5</b>	110M

*Supplement: LARGE models*

XLM-R <sub>LARGE</sub> (Conneau et al., 2019)	85.2	550M
CamemBERT <sub>LARGE</sub> (fine-tuned)	<u>85.7</u>	335M

Table 4: **NLI** accuracy on the French XNLI test set (best model selected on validation out of 10). Best scores in bold, second best underlined.

# Multilingual vs Monolingual models in English - Closer look to Translate TEST

Do we need these huge multilingual LMs?

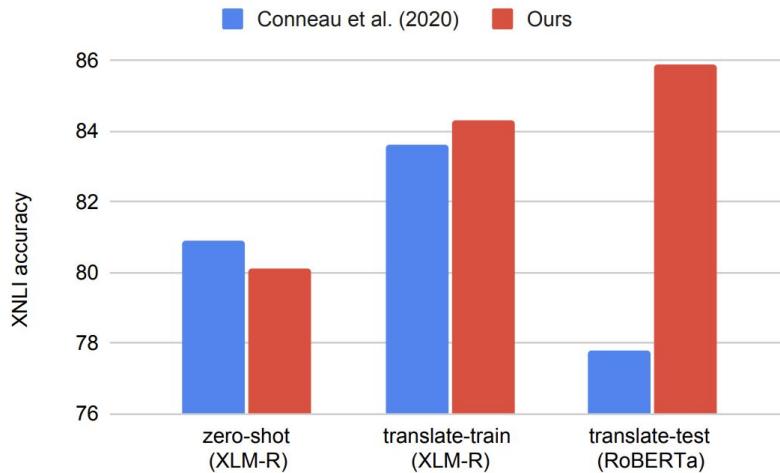


Figure 1: **XNLI accuracy.** We show that *translate-test* can do substantially better than previously reported.

Translate-TEST is highly sensitive to the choice of MT system. With SOTA MT (here NLLB 3.3B), monolingual model outperforms multilingual ones

# Multilingual vs Monolingual models in English - Closer look to Translate TEST

		xnli	pwsx	marc	xcop	xsto	exm	avg
XLM-R	zero-shot	80.1	87.1	60.6	<u>69.1</u>	84.6	<u>36.0</u>	69.6
	translate-train	84.3	<b>90.7</b>	<b>60.8</b>	67.6	86.7	35.2	<u>70.9</u>
	translate-test	vanilla	79.3	86.9	58.0	68.6	84.8	34.9
RoBERTa	translate-test	ours	<u>84.6</u>	89.3	58.8	69.0	<u>87.9</u>	35.0
		vanilla	79.9	87.3	57.6	72.9	89.3	36.3
DeBERTa	translate-test	ours	<u>85.9</u>	<u>89.3</u>	<u>59.1</u>	<u>75.7</u>	<u>91.2</u>	<u>36.4</u>
		vanilla	81.0	87.1	58.2	77.7	92.1	<b>46.1</b>
		ours	<b>86.7</b>	<u>90.3</u>	<u>59.2</u>	<b>81.3</b>	<b>93.8</b>	46.0
								<u>76.2</u>

Table 4: Main results. All systems use NLLB for MT. Best model results underlined, best overall results in **bold**.

# Multilingual vs Monolingual models in English - Closer look to Translate TEST

How to achieve this result?

**Reducing mismatch between training data (human generated) and test data (MT generated) in translate-TEST**

- 1) MT adaptation:
  - Finetune NLLB on training data (in English) with backtranslation (domain transfer)
  - Adapt NLLB to documents by concatenating all sentences
- 2) Training data adaptation:
  - Translate training data into another language + back into English. Then combine with the original training data.

- Neural Machine Translation
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Learning a new language with LLMs
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models
  - **Multicultural NLP**

## Multicultural NLP: Problem statement

LLMs are trained on large-scale web data, however the internet is mainly composed of content related to the western culture. Are LLMs able to reason across different cultures?

First, what is culture and how to measure it?

*Adilazuarda et al. (2024). Towards Measuring and Modeling “Culture” in LLMs: A Survey*

# Multicultural NLP: Measuring Culture in LLMs

*Adilazuarda et al. (2024). Towards Measuring and Modeling “Culture” in LLMs: A Survey*

Hard to define but it exists “*proxies of cultures*”:

1) Demographic proxies

- Culture is, almost always, described at the level of a community or group of people, who share certain common demographic attributes. These could be ethnicity (Masai culture), religion (Islamic culture), age (Gen Z culture), socio-economic class (middle class or urban) etc. + their intersection.

2) Semantic proxies

- Often cultures are defined in terms of the emotions and values, food and drink, kinship terms, social etiquette, etc. prevalent within a group of people.

# Multicultural NLP: Measuring Culture in LLMs

*Adilazuarda et al. (2024). Towards Measuring and Modeling “Culture” in LLMs: A Survey*

**Probing method** mainly used: prompt the LLM and analyze the output (or the logit distribution) = black-box approach

Example:

Pick one.

Do people in [COUNTRY\_NAME] believe that  
claiming government benefits to which you  
are not entitled is:

1. Never justifiable
2. Something in between
3. Always justifiable

# Multicultural NLP: Measuring Culture in LLMs

*Zhong et al. (2024). Cultural Value Differences of LLMs: Prompt, Language, and Model Size*

- (i) LLMs exhibit relatively consistent cultural values when presented with prompts in a single language.
- (ii) The prompting language e.g., Chinese or English, can influence the expression of cultural values. The same question can elicit divergent cultural values when the same LLM is queried in a different language.
- (iii) Differences in sizes of the same model (e.g., Llama2-7B vs 13B vs 70B) have a more significant impact on their demonstrated cultural values than model differences (e.g., Llama2 vs Mixtral). Our experiments reveal that query language and model size of LLM are the main factors resulting in cultural value differences.

# Multicultural NLP: Measuring Culture in LLMs

*Zhong et al. (2024). Cultural Value Differences of LLMs: Prompt, Language, and Model Size*

- (i) LLMs exhibit relatively consistent cultural values when presented with prompts in a single language.
- (ii) The prompting language e.g., Chinese or English, can influence the expression of cultural values. The same question can elicit divergent cultural values when the same LLM is queried in a different language.
- (iii) Differences in sizes of the same model (e.g., Llama2-7B vs 13B vs 70B) have a more significant impact on their demonstrated cultural values than model differences (e.g., Llama2 vs Mixtral). Our experiments reveal that query language and model size of LLM are the main factors resulting in cultural value differences.

# Multicultural NLP: Measuring Culture in LLMs

*Zhong et al. (2024). Cultural Value Differences of LLMs: Prompt, Language, and Model Size*

Measure used => Value Survey Module (VSM)

- Measuring cultural values as outlined in Hofstede's Cultural Dimensions Theory
  - 24 questions to evaluate the interviewees' cultural values, each question has 5 possible answers.

## INTERNATIONAL QUESTIONNAIRE (VSM 2013)- page 1

Please think of an ideal job, disregarding your present job, if you have one. In choosing an ideal job, how important would it be to you to ... (please circle one answer in each line across):

- 1 = of utmost importance
- 2 = very important
- 3 = of moderate importance
- 4 = of little importance
- 5 = of very little or no importance

01. have sufficient time for your personal or home life	1	2	3	4	5
02. have a boss (direct superior) you can respect	1	2	3	4	5
03. get recognition for good performance	1	2	3	4	5
04. have security of employment	1	2	3	4	5
05. have pleasant people to work with	1	2	3	4	5

Scoring system, based on answers, comprising six dimensions for measuring cultural values: Power Distance (**PDI**), Individualism (**IDV**), Uncertainty Avoidance (**UAI**), Masculinity (**MAS**), Long-term Orientation (**LTO**), and Indulgence (**IVR**)

# Multicultural NLP: Measuring Culture in LLMs

*Zhong et al. (2024). Cultural Value Differences of LLMs: Prompt, Language, and Model Size*

LLMs were presented with a curated collection of simulated identities, each comprising three variables: (i) nationality, (ii) age, and (iii) gender to furnish context for the model's responses to questions.

The study encompasses nine nationalities (refer to the full list in Appendix C), two genders, and three age groups (25, 35, 45), resulting in a total of 54 identities. These variables align with the VSM survey, encompassing interviewees from various countries, genders, and ages. The chosen nations are globally diverse, representing a range of cultures.

# Multicultural NLP: Measuring Culture in LLMs

*Zhong et al. (2024). Cultural Value Differences of LLMs: Prompt, Language, and Model Size*

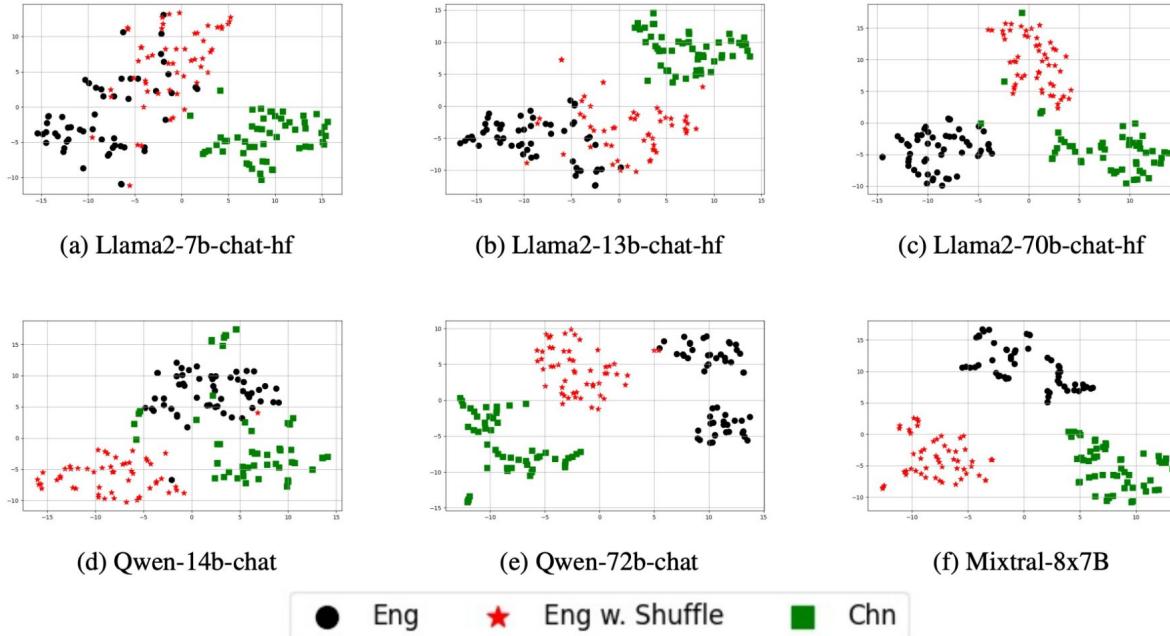


Figure 1: The 6-d VSM scores for different experiment sets for each model are visualized using the t-SNE technique ([van der Maaten and Hinton, 2008](#)) to facilitate direct comparisons. Results from English queries (denoted as "Eng") are displayed with black circles; results from English with Shuffled Options (denoted as "Eng w. Shuffle") are shown with pink stars; and results from Chinese (denoted as "Chn") are represented by green squares.