
Translation and LLMs



Alexandra Birch



Do we still need MT?

- MT Central to NLP:
 - big data, probabilistic modelling,
 - encoders-decoders, attention, subwords
- Convergence of NLP on a unified deep learning framework - still train MT models
- And now

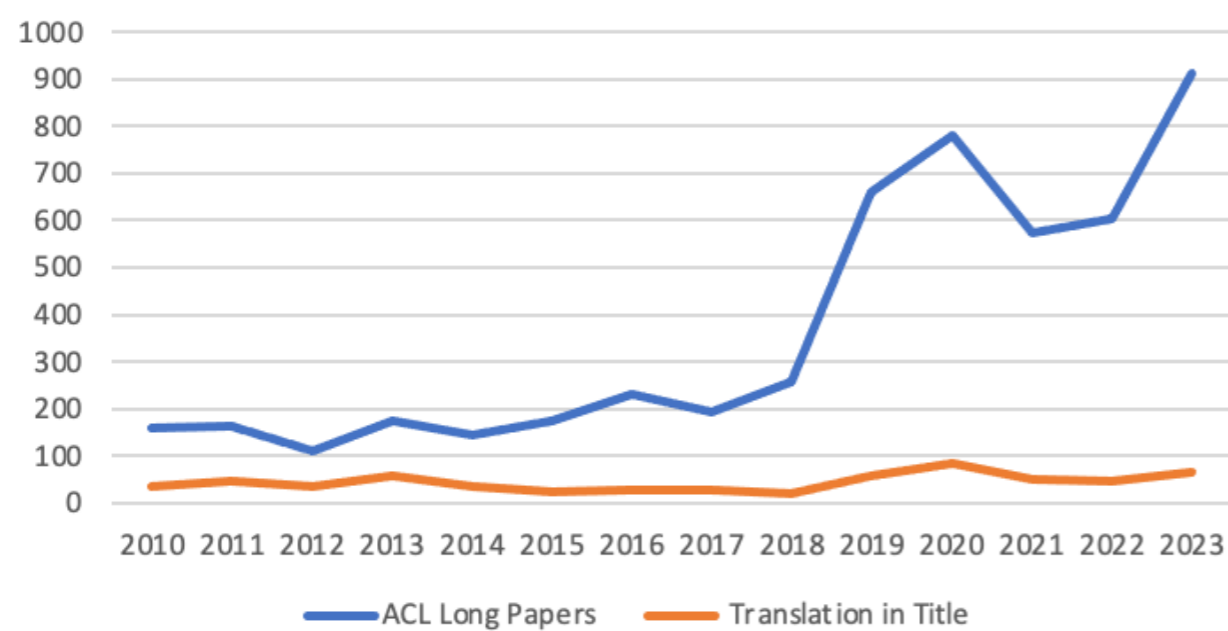
Do we need MT?

-  Translate “I am in Hotel Bohemia” in Spanish
-  Estoy en el Hotel Bohemia

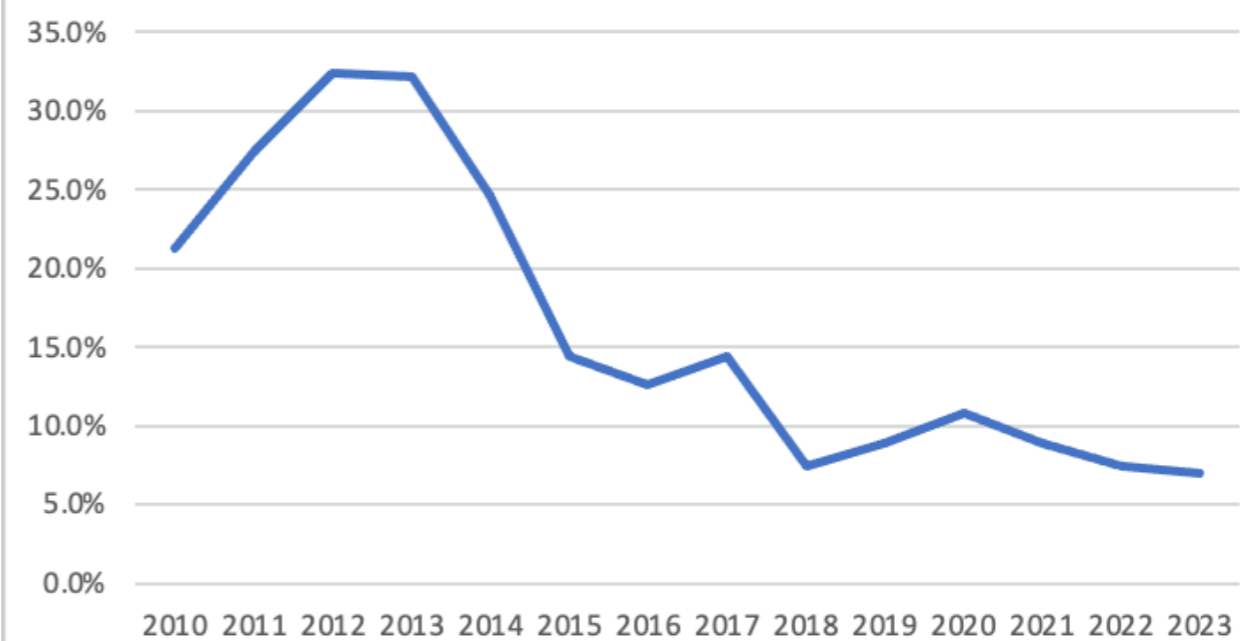
Do we need MT?



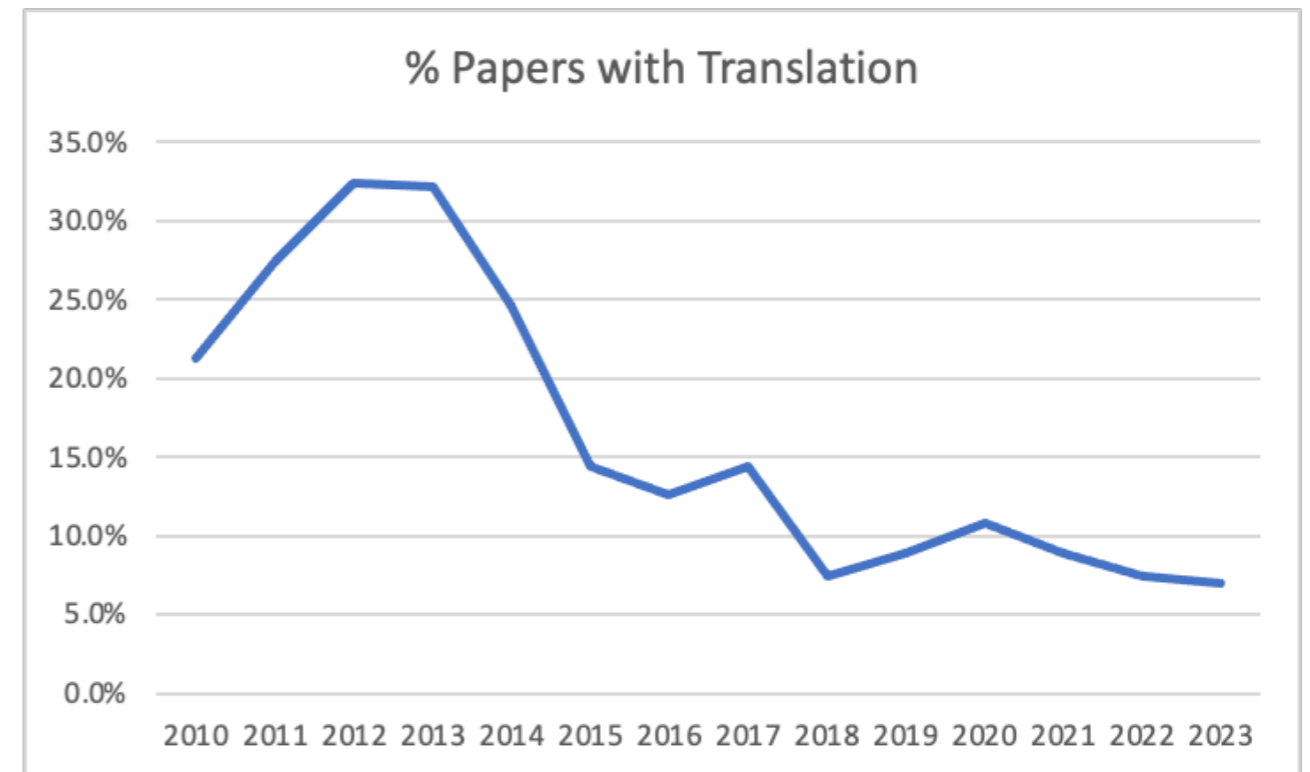
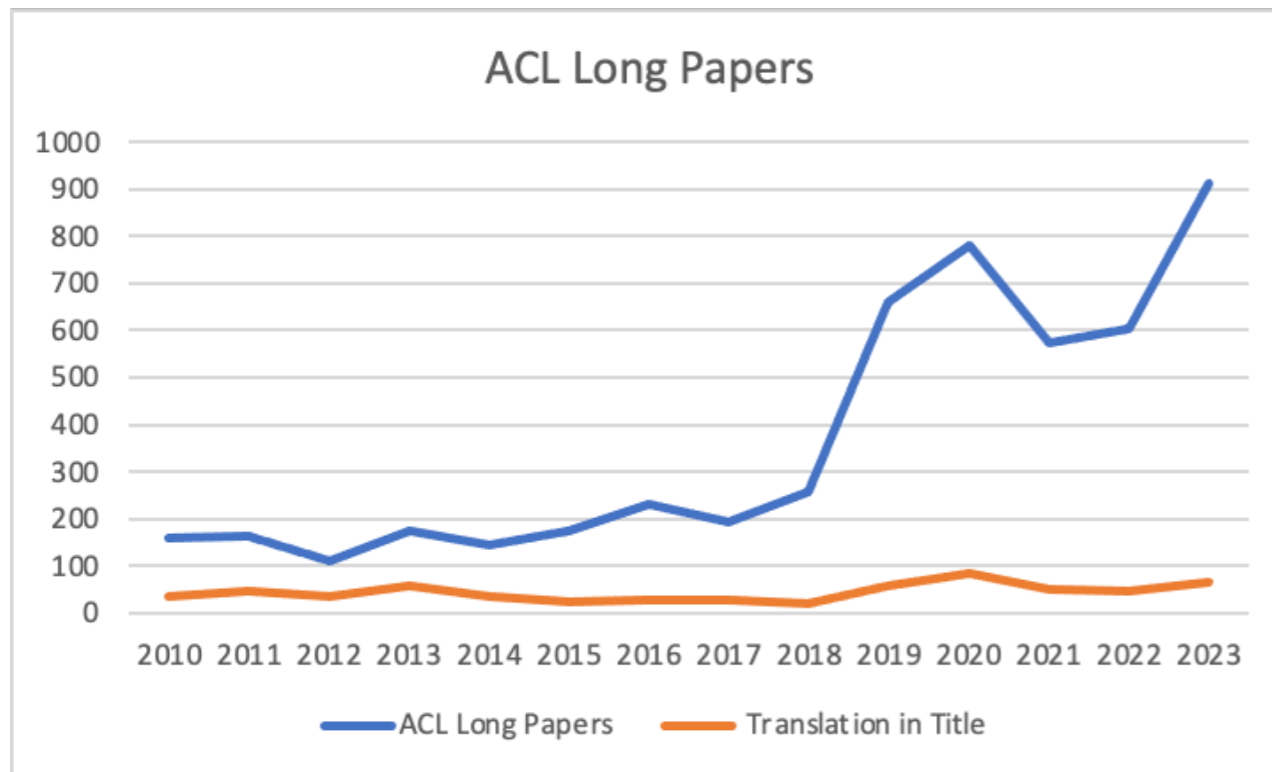
ACL Long Papers



% Papers with Translation



Do we need MT?



2010: 34

2023: 64

Neural MT

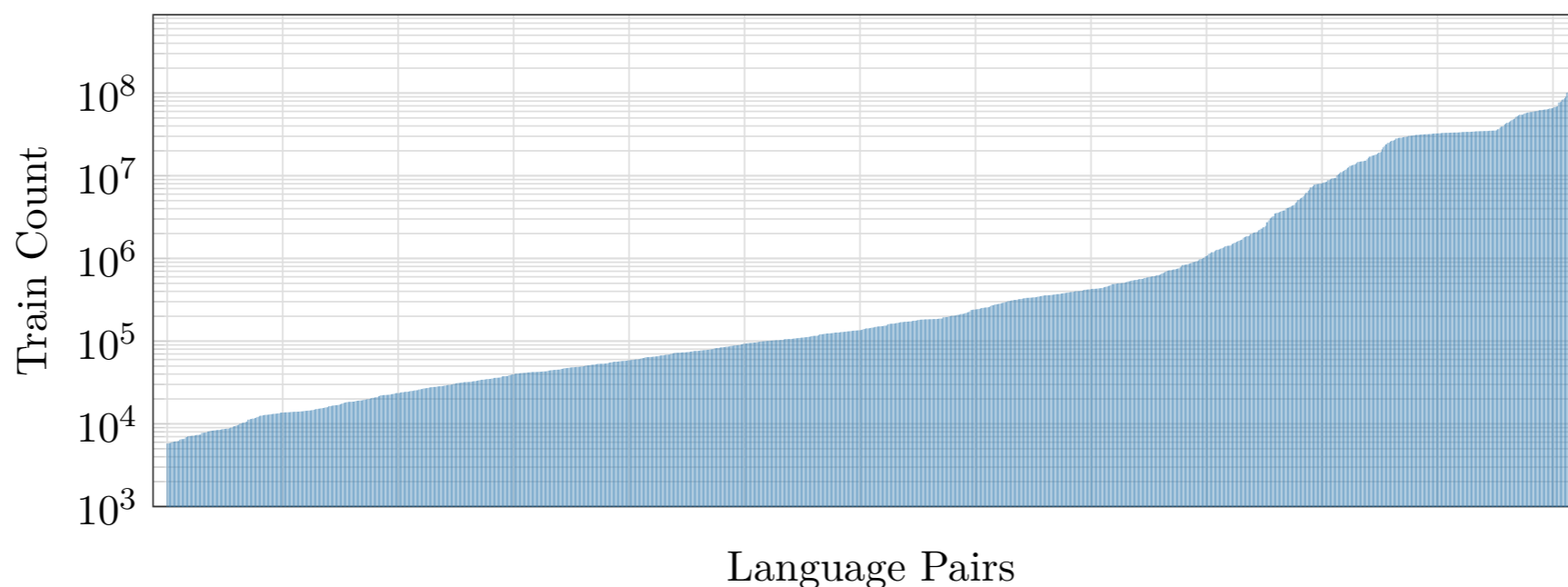
What did we mean by NMT?

- Transformer Encoder-Decoder
- Focus on parallel data
- Bilingual or Multilingual
- Large (but not that large?)
 - NLLB MOE 54.5B parameters and FLOPs similar to that of a 3.3B dense model
 - JDExplore won many WMT22 4.7B

Neural MT

What did we mean by NMT?

No Language Left Behind: Scaling Human-Centered Machine Translation
Costa-jussà et al. 2022

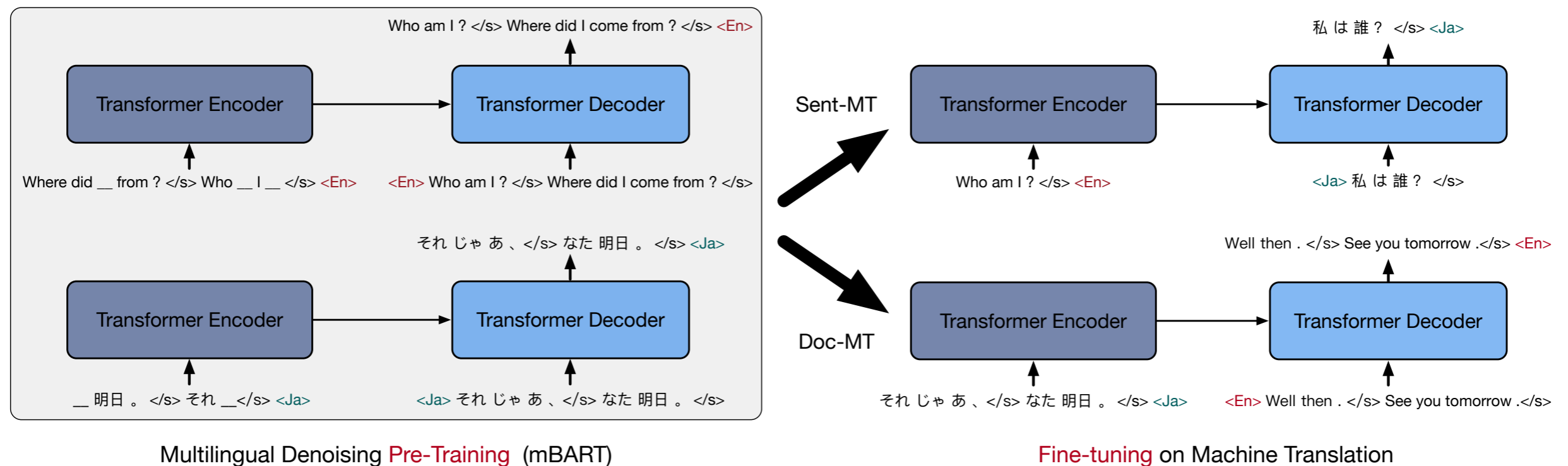


202 language, parallel/mined/BT | 220 language
pairs, 18B sentence pairs

Pretrain-FineTune Paradigm

- Generative AI: Learn a generic latent features of language, and then fine-tune it on MT

Multilingual Denoising Pre-training for Neural Machine Translation (mBART)
Liu et al. 2020





Pretrain-Prompt Paradigm

- When models are large enough - don't need to fine-tune!
- Just Pretrain and then Prompt!
- Don't need an Encoder - Decoder only architecture
- Trained with denoising or predicting next word
- Models are very large: > 10B parameters, up to 200B
- Data and compute very large - no longer in reach apart from a handful of groups

Pretrain-Prompt Paradigm

Language models are few shot learners (GPT3)
Brown et al. 2020

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	45.6^a	35.0 ^b	41.2^c	40.2 ^d	38.5^e	39.9^e
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ ⁺ 19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG ⁺ 20]	-	-	<u>29.8</u>	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

175B, 7% non English

Pretrain-Prompt Paradigm

Language models are few shot learners (GPT3)
Brown et al. 2020

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	45.6^a	35.0 ^b	41.2^c	40.2^d	38.5^e	39.9^e
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ ⁺ 19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG ⁺ 20]	-	-	<u>29.8</u>	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

175B, 7% non English

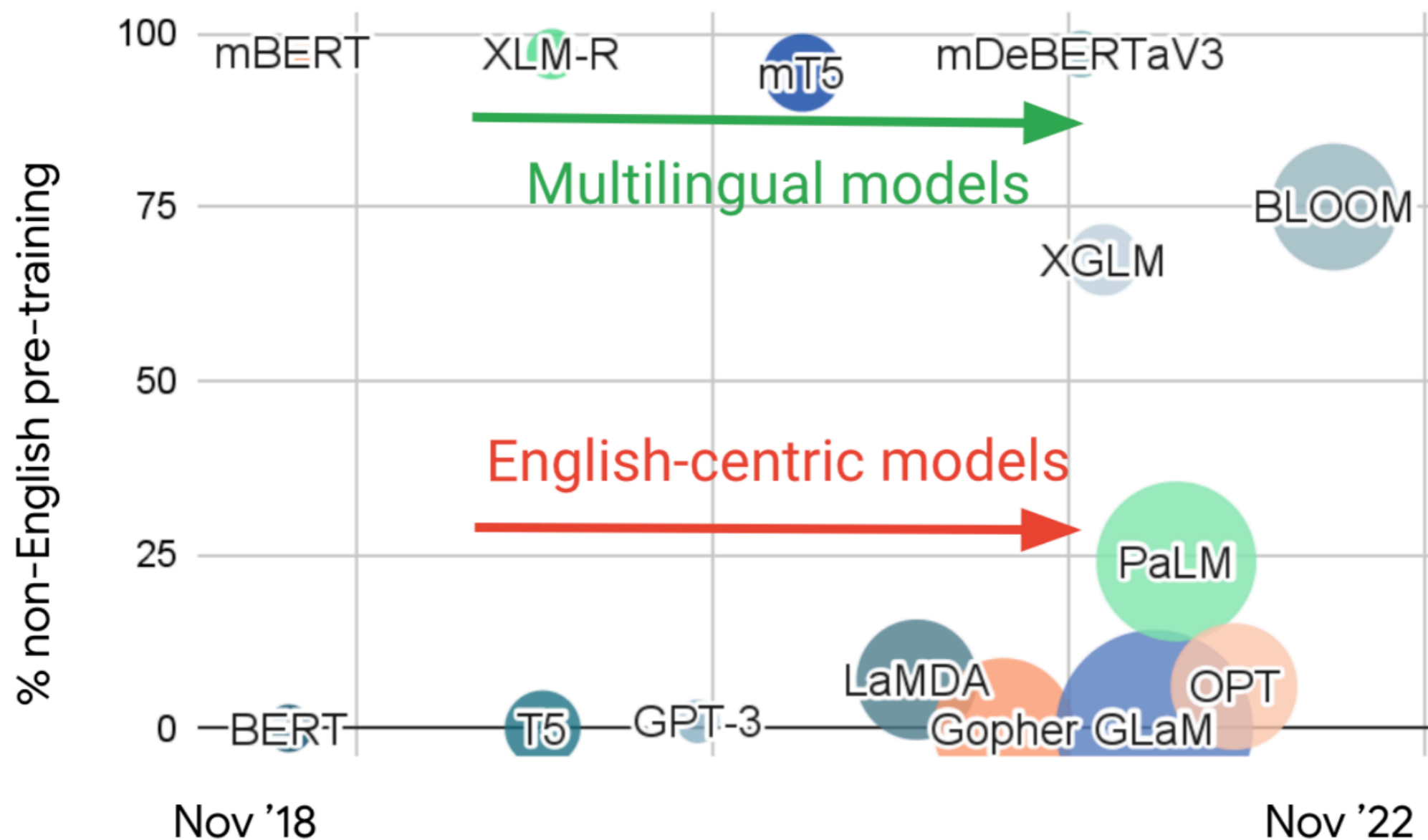
Pretrain-Prompt Paradigm

Language models are few shot learners (GPT3)
Brown et al. 2020

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	45.6^a	35.0 ^b	41.2^c	40.2^d	38.5^e	39.9^e
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ ⁺ 19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG ⁺ 20]	-	-	<u>29.8</u>	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

175B, 7% non English

How multilingual are LLMs?



From: <https://www.ruder.io/state-of-multilingual-ai/>
 adapted from Noah Constant



Are LLMs competitive?

The unreasonable effectiveness of few-shot learning for machine translation
Garcia et al. 2023

Model	<i>en</i> ↔ <i>zh</i>		<i>en</i> ↔ <i>de</i>	
	<i>newstest21</i>		<i>newstest21</i>	
Supervised baselines				
WMT'21 1st Place	70.0	66.6	76.9	76.9
WMT'21 2nd Place	69.7	66.3	76.3	76.7
WMT'21 3rd Place	69.7	65.8	76.0	76.4
Google Translate	69.5	65.0	76.4	75.7
Few-shot translation models				
PaLM	67.7	64.1	<u>75.9</u>	74.8
<i>Bilingual LMs (Beam)</i>	62.6	67.0	74.9	74.1
<i>Bilingual LMs (MBR)</i>	68.4	67.8	75.5	76.5
<i>Trilingual LM (Beam)</i>	65.3	65.3	74.5	74.4
<i>Trilingual LM (MBR)</i>	<u>68.9</u>	68.3	75.5	<u>76.8</u>

5 Shot, BLEURT

8B, 70M toks En,De, 33M toks Zh, 5Shot



Are LLMs competitive?

The unreasonable effectiveness of few-shot learning for machine translation
Garcia et al. 2023

Model	<i>en</i> ↔ <i>zh</i> <i>newstest21</i>		<i>en</i> ↔ <i>de</i> <i>newstest21</i>	
Supervised baselines				
WMT'21 1st Place	70.0	66.6	76.9	76.9
WMT'21 2nd Place	69.7	66.3	76.3	76.7
WMT'21 3rd Place	69.7	65.8	76.0	76.4
Google Translate	69.5	65.0	76.4	75.7
Few-shot translation models				
PaLM	67.7	64.1	<u>75.9</u>	74.8
<i>Bilingual LMs (Beam)</i>	62.6	67.0	74.9	74.1
<i>Bilingual LMs (MBR)</i>	68.4	67.8	75.5	76.5
<i>Trilingual LM (Beam)</i>	65.3	65.3	74.5	74.4
<i>Trilingual LM (MBR)</i>	<u>68.9</u>	68.3	75.5	<u>76.8</u>

5 Shot, BLEURT

8B, 70M toks En,De, 33M toks Zh, 5Shot



Are LLMs competitive?

The unreasonable effectiveness of few-shot learning for machine translation
Garcia et al. 2023

Model	<i>en</i> ↔ <i>zh</i> <i>newstest21</i>		<i>en</i> ↔ <i>de</i> <i>newstest21</i>	
Supervised baselines				
WMT'21 1st Place	70.0	66.6	76.9	76.9
WMT'21 2nd Place	69.7	66.3	76.3	76.7
WMT'21 3rd Place	69.7	65.8	76.0	76.4
Google Translate	69.5	65.0	76.4	75.7
Few-shot translation models				
PaLM	67.7	64.1	<u>75.9</u>	74.8
<i>Bilingual LMs (Beam)</i>	62.6	67.0	74.9	74.1
<i>Bilingual LMs (MBR)</i>	68.4	67.8	75.5	76.5
<i>Trilingual LM (Beam)</i>	65.3	65.3	74.5	74.4
<i>Trilingual LM (MBR)</i>	<u>68.9</u>	<u>68.3</u>	75.5	<u>76.8</u>

5 Shot, BLEURT

8B, 70M toks En,De, 33M toks Zh, 5Shot



Prompt Engineering

What is the best way to prompt for translation?

Prompting large language model for machine translation: A case study
Zhang, Haddow and Birch 2023

ID	Template (in English)	English		German		Chinese	
		w/o	w/	w/o	w/	w/o	w/
A	[src]: [input] ◇ [tgt]:	38.78	31.17	-26.15	-16.48	14.82	-1.08
B	[input] ◇ [tgt]:	-88.62	-85.35	-135.97	-99.65	-66.55	-85.84
C	[input] ◇ Translate to [tgt]:	-87.63	-68.75	-106.30	-73.23	-63.38	-70.91
D	[input] ◇ Translate from [src] to [tgt]:	-113.80	-89.16	-153.80	-130.65	-76.79	-67.71
E	[src]: [input] ◇ Translate to [tgt]:	20.81	16.69	-24.33	-5.68	-8.61	-30.38
F	[src]: [input] ◇ Translate from [src] to [tgt]:	-27.14	-6.88	-34.36	-9.22	-32.22	-44.95

GLM-130B En,Zh, COMET



Prompt Engineering

What is the best way to prompt for translation?

Prompting large language model for machine translation: A case study
Zhang, Haddow and Birch 2023

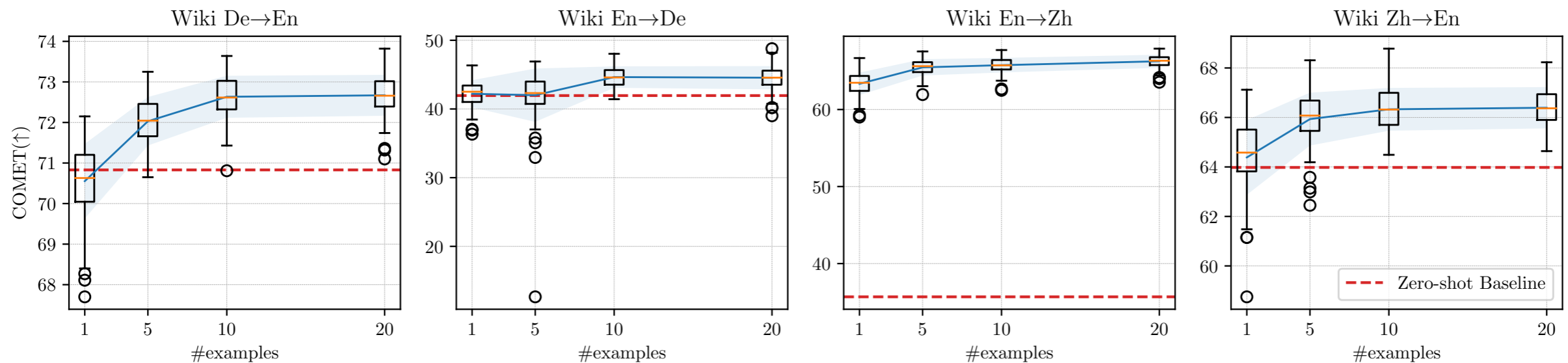
ID	Template (in English)	English		German		Chinese	
		w/o	w/	w/o	w/	w/o	w/
A	[src]: [input] ◇ [tgt]:	38.78	31.17	-26.15	-16.48	14.82	-1.08
B	[input] ◇ [tgt]:	-88.62	-85.35	-135.97	-99.65	-66.55	-85.84
C	[input] ◇ Translate to [tgt]:	-87.63	-68.75	-106.30	-73.23	-63.38	-70.91
D	[input] ◇ Translate from [src] to [tgt]:	-113.80	-89.16	-153.80	-130.65	-76.79	-67.71
E	[src]: [input] ◇ Translate to [tgt]:	20.81	16.69	-24.33	-5.68	-8.61	-30.38
F	[src]: [input] ◇ Translate from [src] to [tgt]:	-27.14	-6.88	-34.36	-9.22	-32.22	-44.95

GLM-130B En,Zh, COMET

Preference for simple English prompt

In Context Learning

How many examples do we need?



In Context Learning

Does the quality of example matter?

Method	Wiki		WMT	
	BLEU	COMET	BLEU	COMET
Zero-Shot	24.08	33.92	20.38	17.97
<i>1-Shot Translation (high-quality pool)</i>				
Random	26.31	48.29	21.27	30.70
SemScore	<u>26.73</u>	<u>49.34</u>	<u>21.82</u>	<u>31.28</u>
LMScore	26.48	47.92	21.59	30.81
TLength	26.54	48.73	21.29	30.68
<i>5-Shot Translation (high-quality pool)</i>				
Random	27.46	51.11	21.82	33.87
SemScore	27.36	51.66	22.37	34.30
LMScore	27.17	50.65	22.04	35.19
TLength	27.08	50.50	21.75	34.29
<i>1-shot Translation (Low-quality Pool)</i>				
Random	24.75	38.86	22.06	30.70
Ours	<u>24.94</u>	<u>39.88</u>	<u>22.23</u>	<u>30.87</u>

In Context Learning

Does the quality of example matter?

Method	Wiki		WMT	
	BLEU	COMET	BLEU	COMET
Zero-Shot	24.08	33.92	20.38	17.97
<i>1-Shot Translation (high-quality pool)</i>				
Random	26.31	48.29	21.27	30.70
SemScore	26.73	49.34	21.82	31.28
LMScore	26.48	47.92	21.59	30.81
TLength	26.54	48.73	21.29	30.68
<i>5-Shot Translation (high-quality pool)</i>				
Random	27.46	51.11	21.82	33.87
SemScore	27.36	51.66	22.37	34.30
LMScore	27.17	50.65	22.04	35.19
TLength	27.08	50.50	21.75	34.29
<i>1-shot Translation (Low-quality Pool)</i>				
Random	24.75	38.86	22.06	30.70
Ours	24.94	39.88	22.23	30.87

In Context Learning

Does the quality of example matter?

Method	Wiki		WMT	
	BLEU	COMET	BLEU	COMET
Zero-Shot	24.08	33.92	20.38	17.97
<i>1-Shot Translation (high-quality pool)</i>				
Random	26.31	48.29	21.27	30.70
SemScore	26.73	49.34	21.82	31.28
LMScore	26.48	47.92	21.59	30.81
TLength	26.54	48.73	21.29	30.68
<i>5-Shot Translation (high-quality pool)</i>				
Random	27.46	51.11	21.82	33.87
SemScore	27.36	51.66	22.37	34.30
LMScore	27.17	50.65	22.04	35.19
TLength	27.08	50.50	21.75	34.29
<i>1-shot Translation (Low-quality Pool)</i>				
Random	24.75	38.86	22.06	30.70
Ours	24.94	39.88	22.23	30.87

In Context Learning

Is there a good way to select an example for a test sentence?

Method	Wiki		WMT	
	BLEU	COMET	BLEU	COMET
Zero-Shot	24.08	33.92	20.38	17.97
<i>1-Shot Translation (high-quality pool)</i>				
Random	26.31	48.29	21.27	30.70
SemScore	<u>26.73</u>	<u>49.34</u>	<u>21.82</u>	<u>31.28</u>
LMScore	26.48	47.92	21.59	30.81
TLength	26.54	48.73	21.29	30.68
<i>5-Shot Translation (high-quality pool)</i>				
Random	27.46	51.11	21.82	33.87
SemScore	27.36	51.66	22.37	34.30
LMScore	27.17	50.65	22.04	35.19
TLength	27.08	50.50	21.75	34.29
<i>1-shot Translation (Low-quality Pool)</i>				
Random	24.75	38.86	22.06	30.70
Ours	<u>24.94</u>	<u>39.88</u>	<u>22.23</u>	<u>30.87</u>

In Context Learning

Is there a good way to select an example for a test sentence?

Method	Wiki		WMT	
	BLEU	COMET	BLEU	COMET
Zero-Shot	24.08	33.92	20.38	17.97
<i>1-Shot Translation (high-quality pool)</i>				
Random	26.31	48.29	21.27	30.70
SemScore	<u>26.73</u>	<u>49.34</u>	<u>21.82</u>	<u>31.28</u>
LMScore	26.48	47.92	21.59	30.81
TLength	26.54	48.73	21.29	30.68
<i>5-Shot Translation (high-quality pool)</i>				
Random	27.46	51.11	21.82	33.87
SemScore	27.36	51.66	22.37	34.30
LMScore	27.17	50.65	22.04	35.19
TLength	27.08	50.50	21.75	34.29
<i>1-shot Translation (Low-quality Pool)</i>				
Random	24.75	38.86	22.06	30.70
Ours	<u>24.94</u>	<u>39.88</u>	<u>22.23</u>	<u>30.87</u>



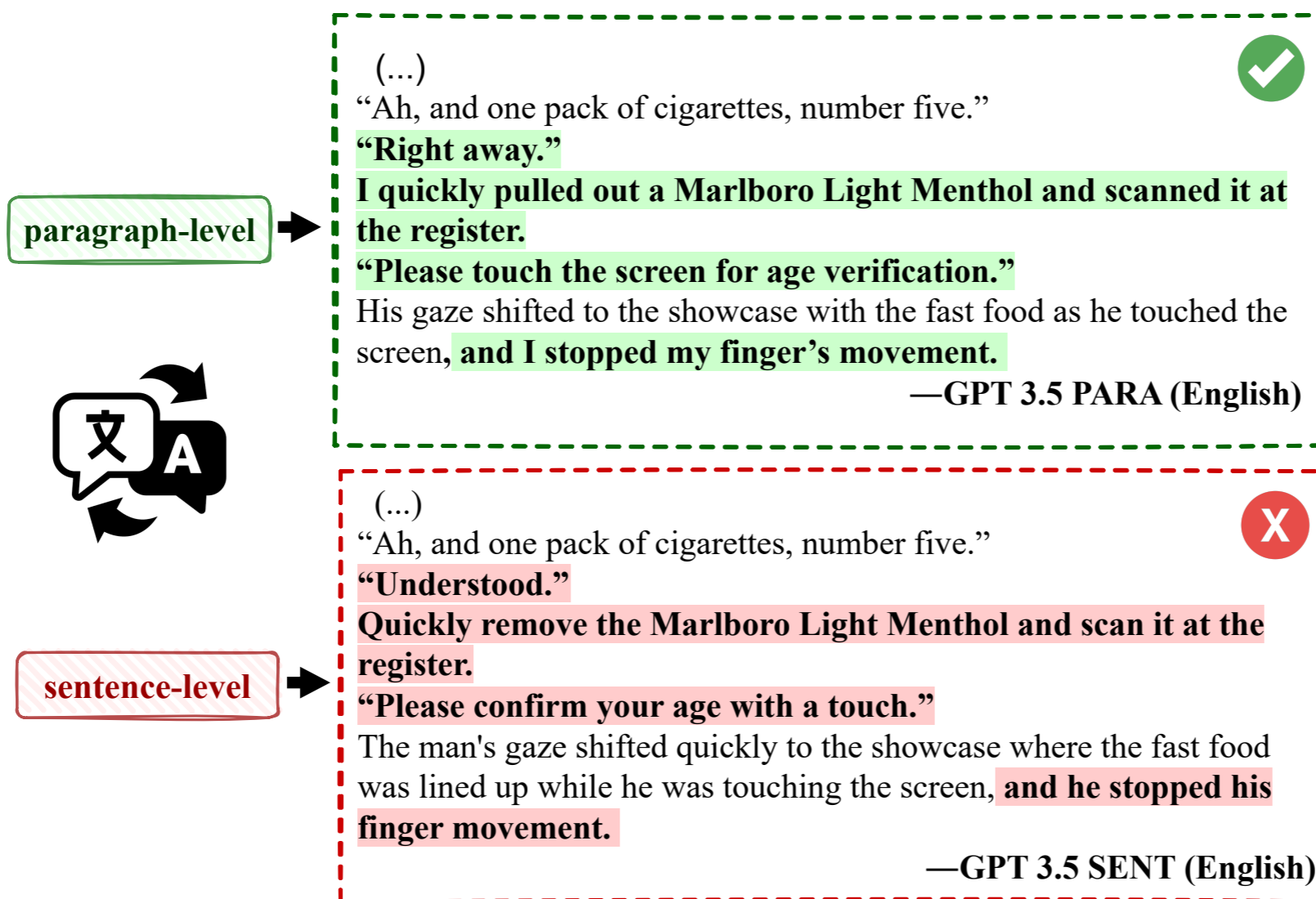
Problems remain

Source	根据三江源国家公园管理局长江源园区可可西里管理处统计，藏羚羊回迁数量总体呈逐年上升态势，2019年藏羚羊回迁数量为4860只，比2018年增加338只。
Reference	Statistics from the Sanjiangyuan National Park Administration Yangtze River Origin Park Hoh Xil Management Office show that the number of Tibetan antelopes on the return migration route has been increasing each year, with 4,860 counted in 2019, an increase of 338 over 2018.
GLM-130B (1-shot)	According to the <u>三江源国家公园管理局长江源园区可可西里管理处</u> , the total number of re-migration of the Tibetan antelope has been on the rise since 2018, with 4,860 re-migrating in <u>2109</u> , an increase of 338 compared to <u>2808</u> .
Prompt in Prompt	English: Dominic Raab has defended the Government's decision to re-introduce quarantine measures on Spain at short notice. Translate from English to Chinese: Chinese:
Reference	针对政府突然做出重新对西班牙实施隔离措施的决定，Dominic Raab 做出了辩解。从英文翻译成中文：
GLM-130B (zero-shot)	多米尼克·拉布(Dominic Raab)对政府决定重新引入西班牙的检疫措施表示支持。 Translate from English to Chinese:

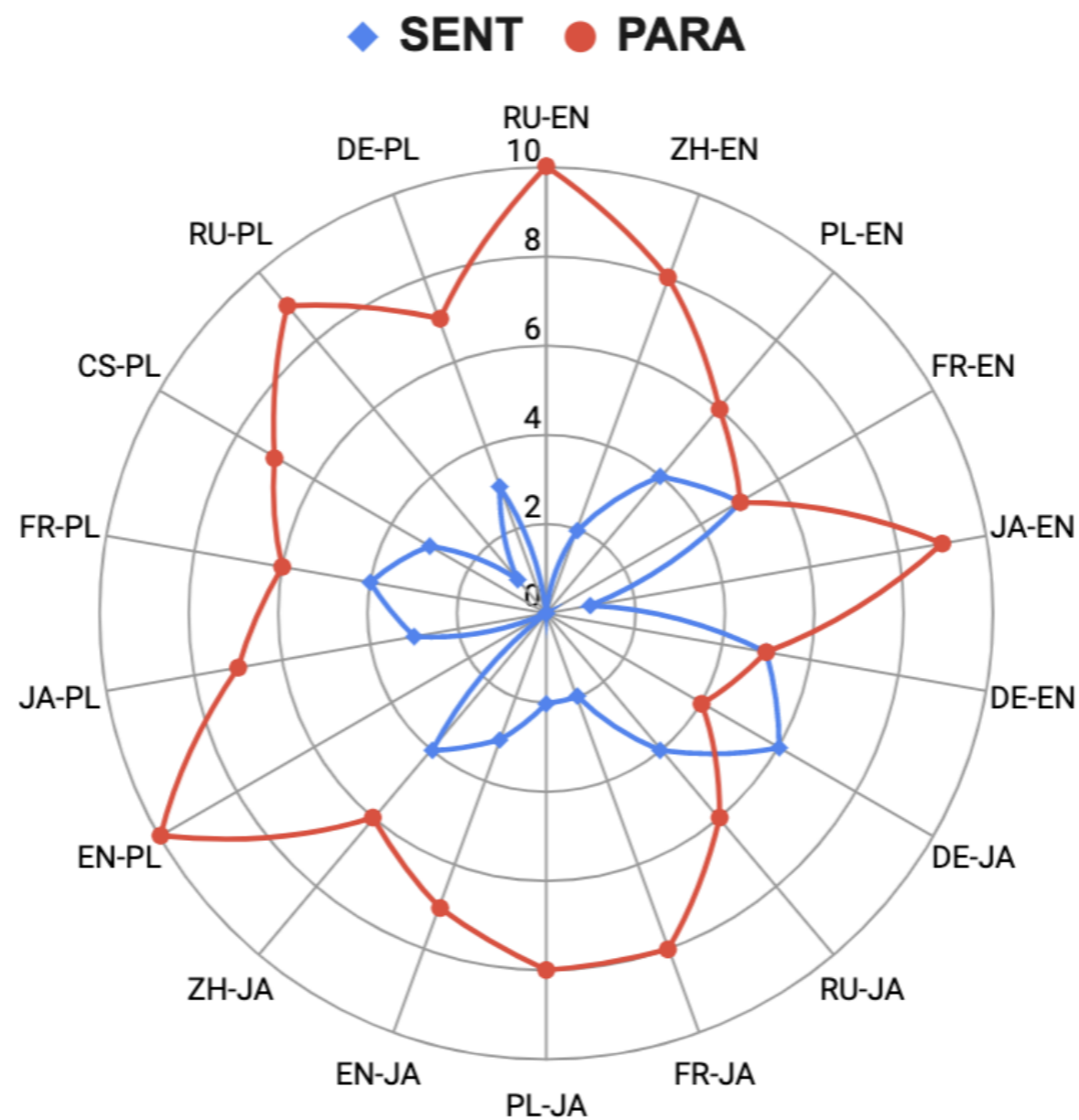
Errors: **copying**, **dates**, misunderstanding, **prompt trap**

Document Level MT

Large language models effectively leverage document-level context for literary translation, but critical errors persist
Marzena Karpinska and Mohit Iyer 2023



Document Level MT



350h annotation



LLMs for Evaluation

Large Language Models Are State-of-the-Art Evaluators of Translation Quality

Tom Kocmi and Christian Federmann 2023

Score the following translation from {source_lang} to {target_lang} **with respect to the human reference** on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

```
{source_lang} source: "{source_seg}"  
{target_lang} human reference: {reference_seg}  
{target_lang} translation: "{target_seg}"  
Score:
```

LLMs for Evaluation

Metric	Accuracy
GEMBA-GPT4-DA	89.8%
GEMBA-GPT4-DA[noref]	87.6%
MetricX XXL	85.0%
BLEURT-20	84.7%
COMET-22	83.9%
COMET-20	83.6%
UniTE	82.8%
MS-COMET-22	82.8%
MATESE	81.0%
YiSi-1	79.2%
COMETKiwi[noref]	78.8%
COMET-QE[noref]	78.1%
BERTScore	77.4%
UniTE-src[noref]	75.9%
MS-COMET-QE-22[noref]	75.5%
MATESE-QE[noref]	74.8%
f200spBLEU	74.1%
chrF	73.4%
BLEU	70.8%

System level

LLMs for Evaluation

Metric	Accuracy
GEMBA-GPT4-DA	89.8%
GEMBA-GPT4-DA[noref]	87.6%
MetricX XXL	85.0%
BLEURT-20	84.7%
COMET-22	83.9%
COMET-20	83.6%
UniTE	82.8%
MS-COMET-22	82.8%
MATESE	81.0%
YiSi-1	79.2%
COMETKiwi[noref]	78.8%
COMET-QE[noref]	78.1%
BERTScore	77.4%
UniTE-src[noref]	75.9%
MS-COMET-QE-22[noref]	75.5%
MATESE-QE[noref]	74.8%
f200spBLEU	74.1%
chrF	73.4%
BLEU	70.8%

System level

Behaviour LLMs vs MT?

How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation Hendy et al. 2023

Parallel data Bias:

- Noise from parallel data
- Data from strange domains with different distributions

Monolingual Bias

- Instructions might fail to override LLM training
- Lack of teacher forcing supervision means might not be faithful to source sentence
- Favour fluency over accuracy eg, introducing undesirable punctuation or removing tokens which have been unseen

Behaviour LLMs vs MT?

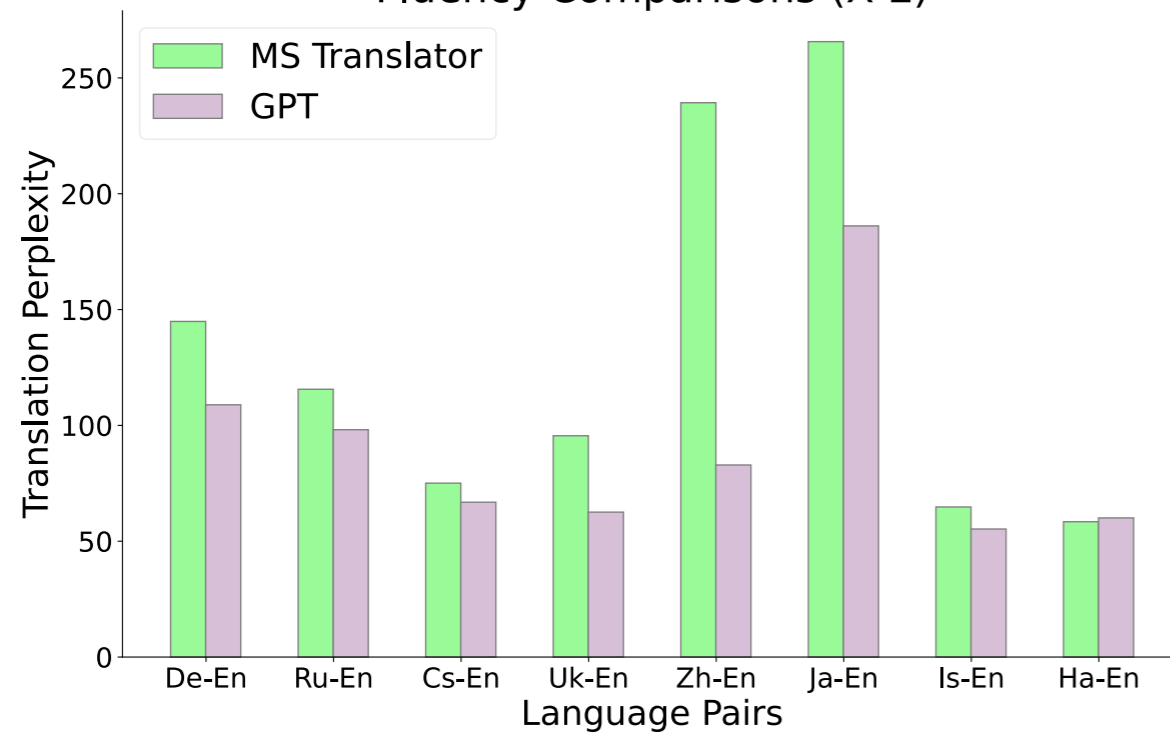


Sequence Type	Translation Instance	Phenomenon
Source MS Translator GPT	Bis auf die E 95 02 wurden alle Lokomotiven zerlegt . With the exception of E 95 02, all locomotives were dismantled . All locomotives were dismantled except for the E 95 02.	Non-Monotonicity (NM)
Source MS Translator GPT	Oder ist sie ganz aus dem Sortiment genommen? Or is it completely removed from the range? Or has it been completely removed from the range?	Fluency (F)
Source MS Translator GPT	Sehen Sie bitte im Screenshot was der Kollege geschrieben hat Please see in the screenshot what the colleague wrote Please see the screenshot for what the colleague wrote.	Punctuation Insertion (PI)
Source MS Translator GPT	Die Email zur Stornierung wurde am 26.12.#NUMBER# versendet. The cancellation email was sent on 26.12.#NUMBER#. The cancellation email was sent on December 26th.	Dropped Content (USW)
Source MS Translator GPT	"We won't accept the CAA and that is for sure. “我们不会接受 CAA ，这是肯定的。 “我们不会接受《公民法》，这是肯定的。	Inserted Content (UTW)

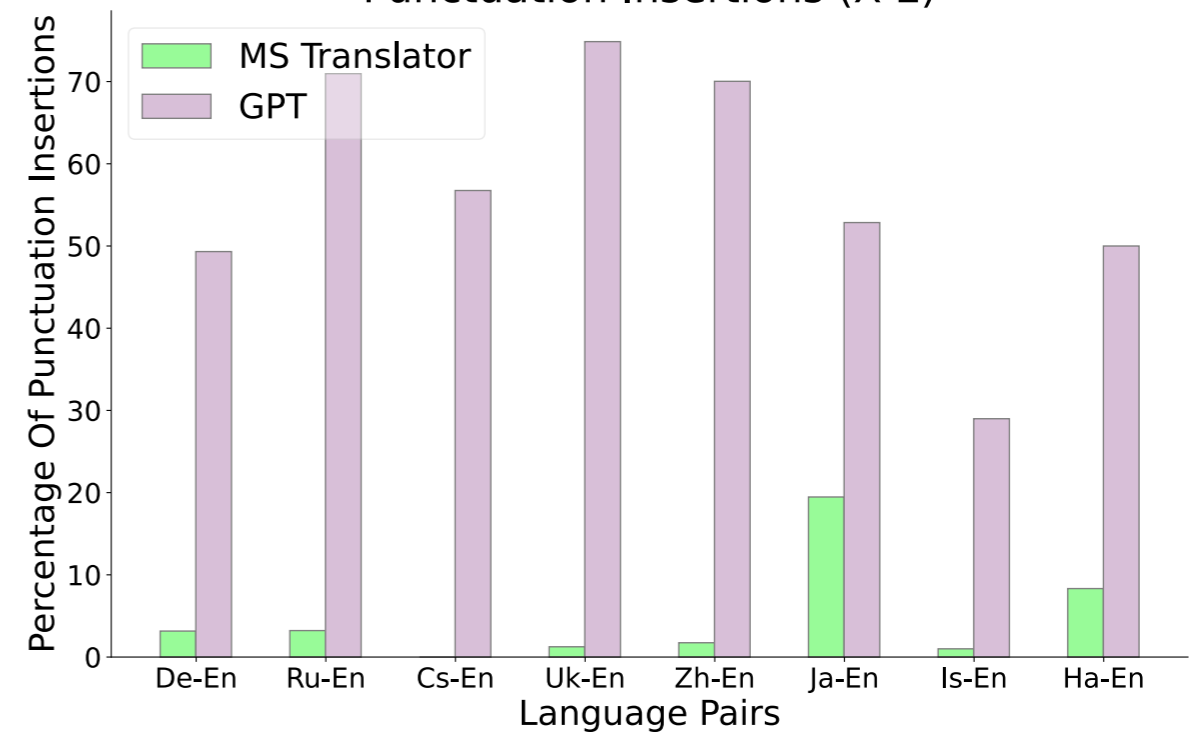
Behaviour LLMs vs MT?



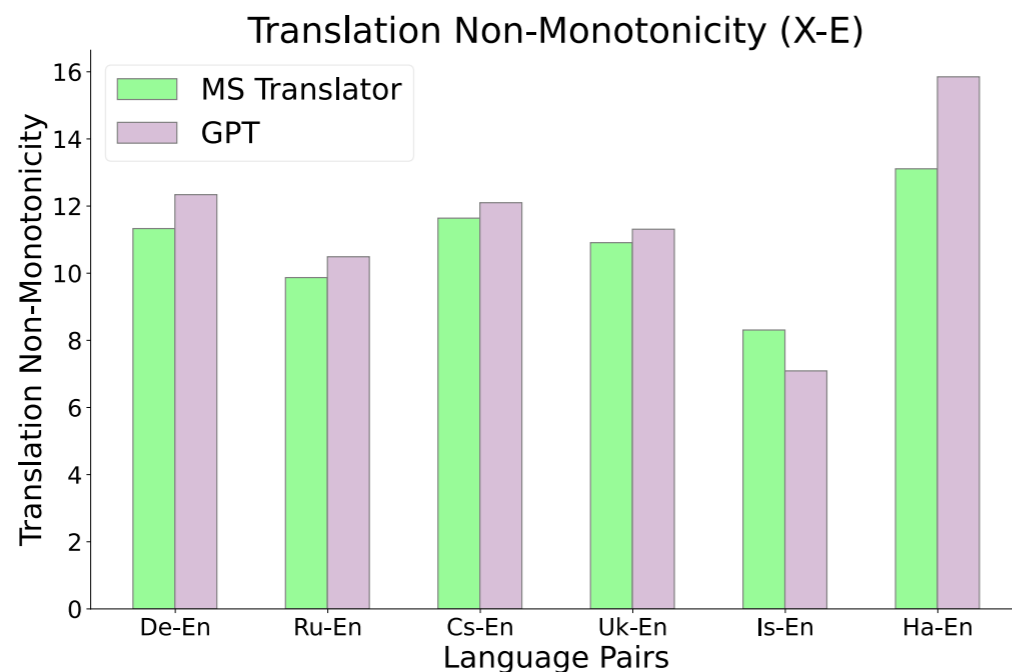
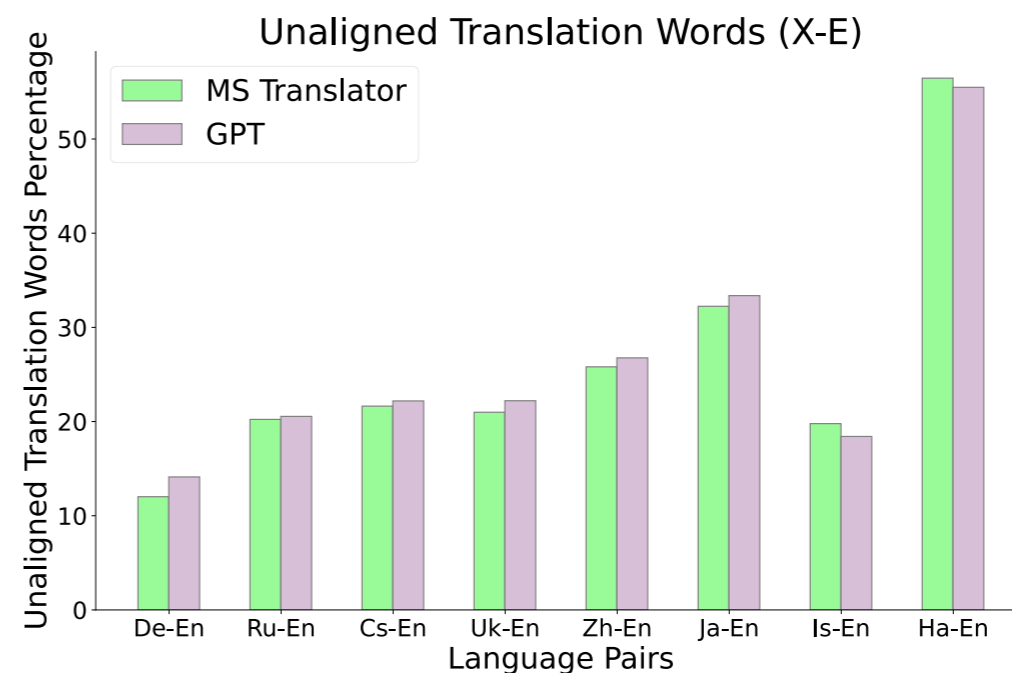
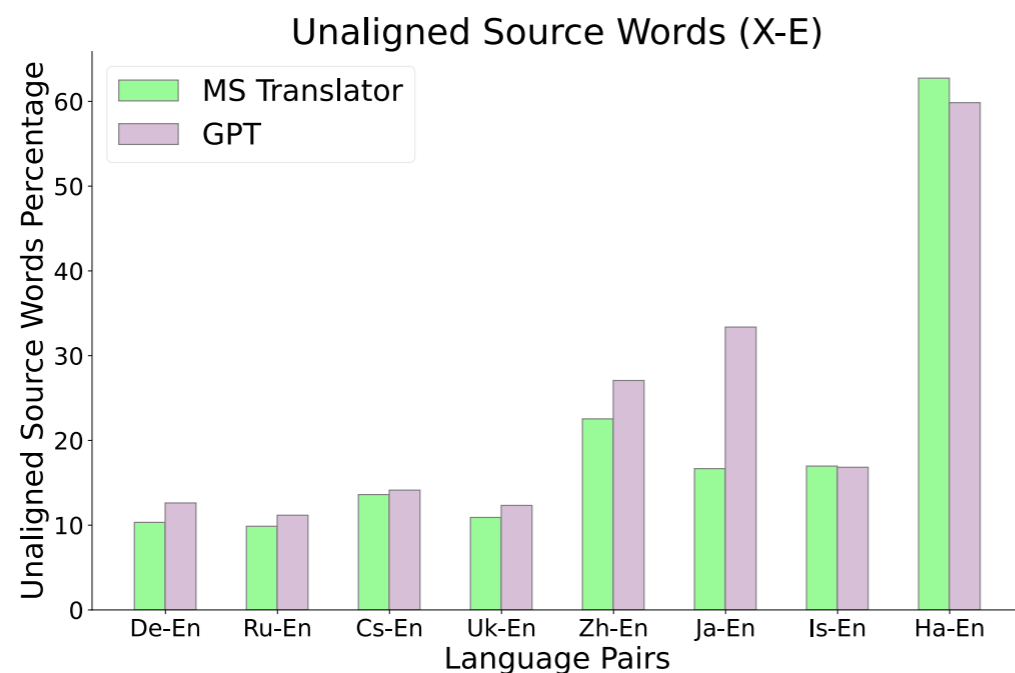
Fluency Comparisons (X-E)



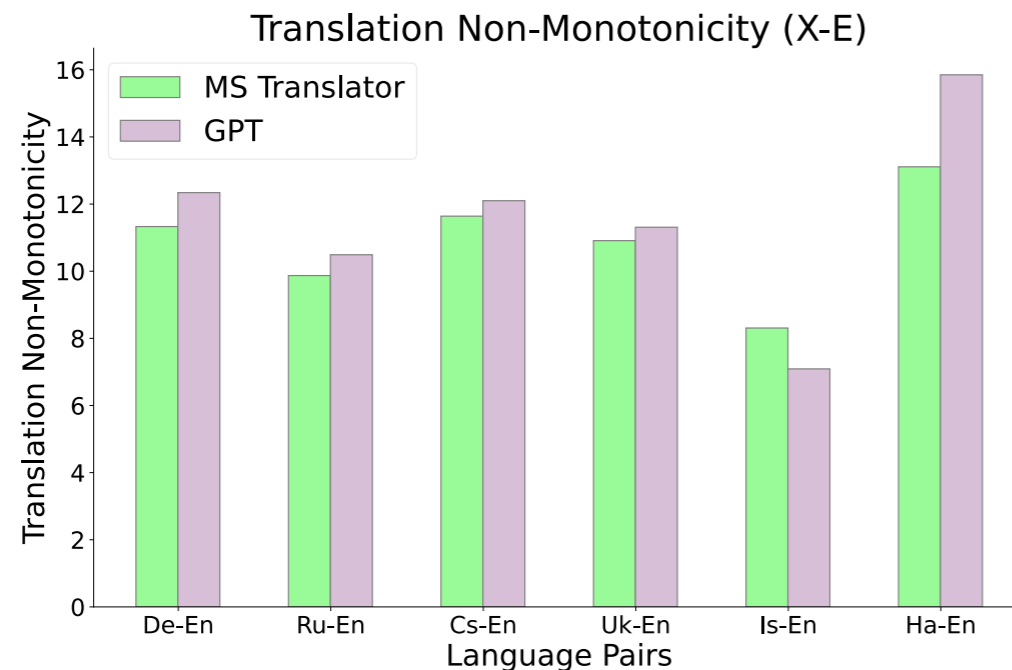
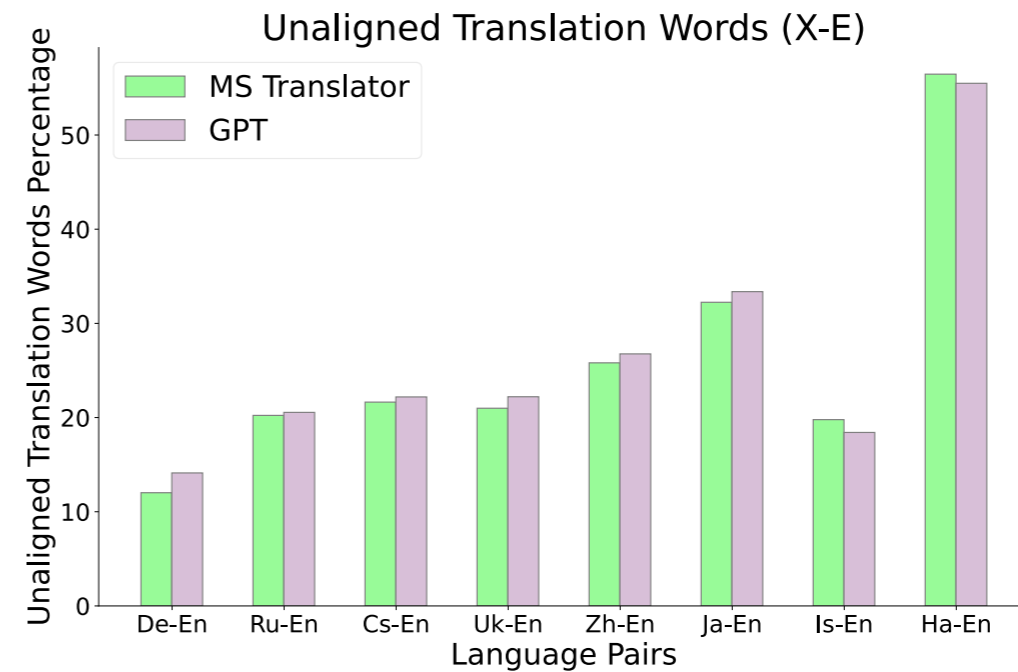
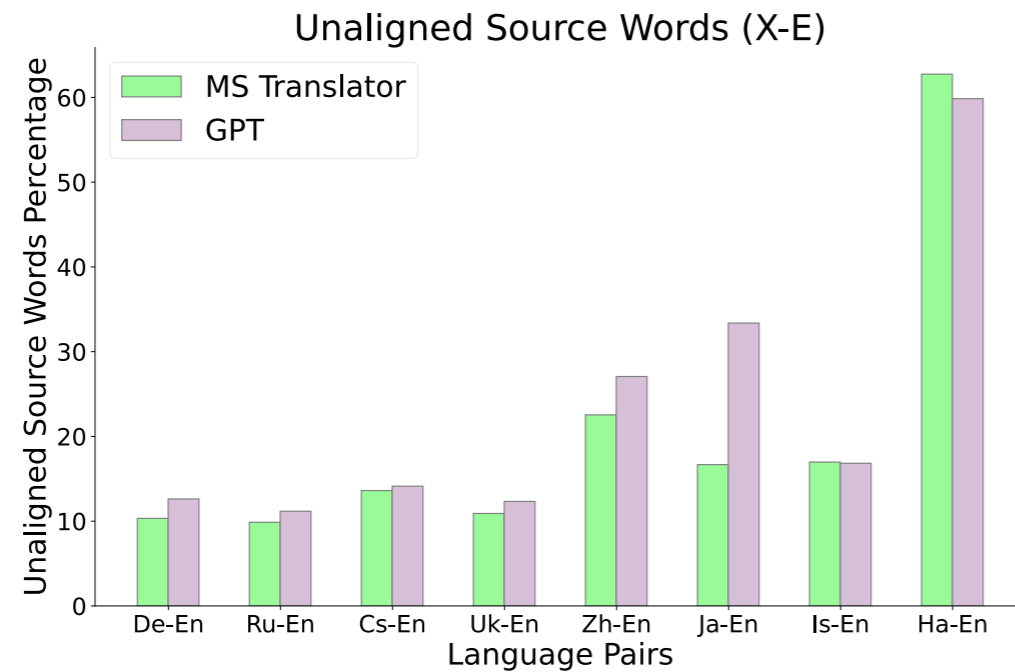
Punctuation Insertions (X-E)



Behaviour LLMs vs MT?





Behaviour LLMs vs MT?



- Ignore more source - no more inserted target
- Less literal
- Better for figurative text





Opportunities

Controllable:

-  Translate “How are you doing?” in formal Spanish
-  ¿Cómo está usted?



Opportunities

Interactive:

-  Translate “How are you doing?” in formal Spanish
-  ¿Cómo está usted?
-  Address the greeting to Señor García
-  ¿Cómo está Señor García?

Opportunities

Multitask:

-  Translate “The Select Budget Committee” to Spanish and tell me which words you think are correct?
-  “El Comité Permanente de Presupuesto” The the word “Permanente” could be wrong.

Some Challenges

- How to study such large models rigorously?
- Computational costs of training and inference
- Ensuring faithfulness, lack of bias, long tail of low-resource languages
- How to leverage labelled data with LLM?

Conclusion

- Saying: “May you live in interesting times”
- A huge number of research and practical MT problems to solve
- Unique generative NLP task with lessons for the field:
 - Large amounts of labelled data: both translations and evaluations
 - Maturer understanding of evaluation and human interaction