Translation and LLMs

Alexandra Birch

Do we still need MT?

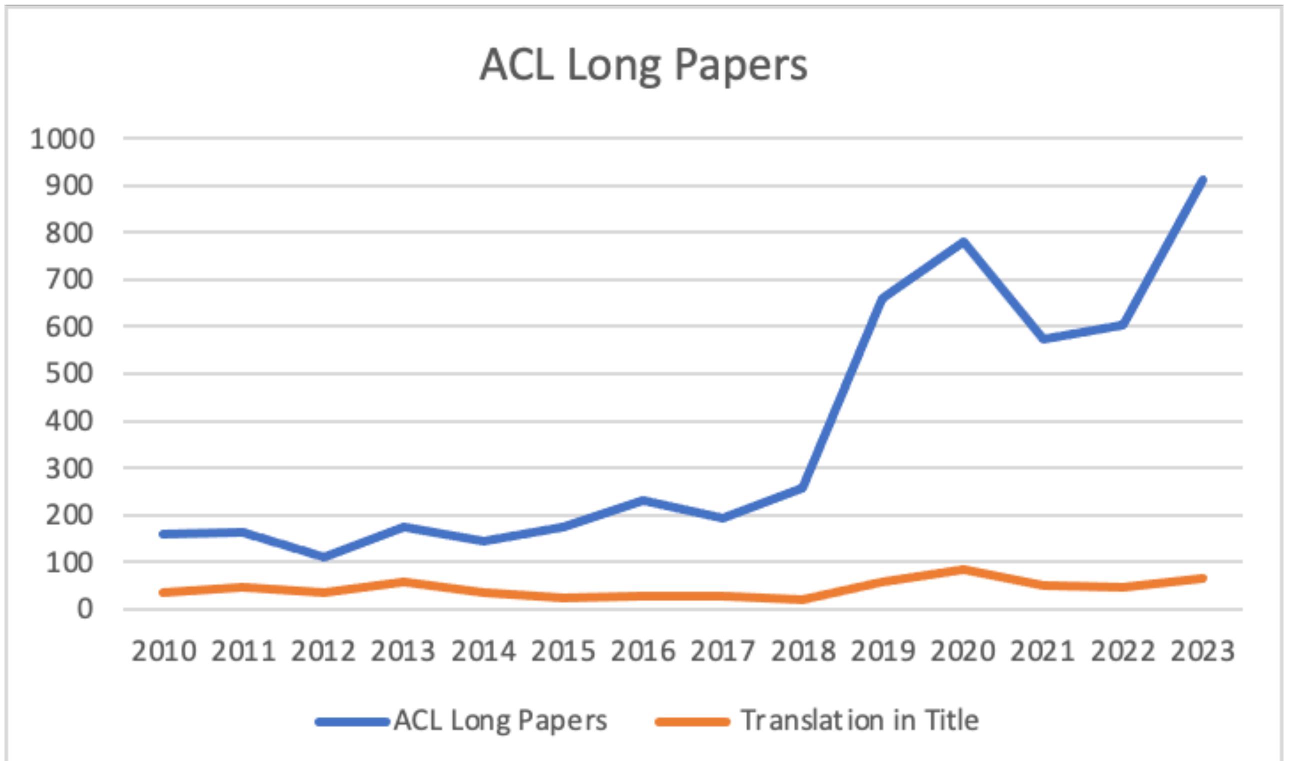
- MT Central to NLP:

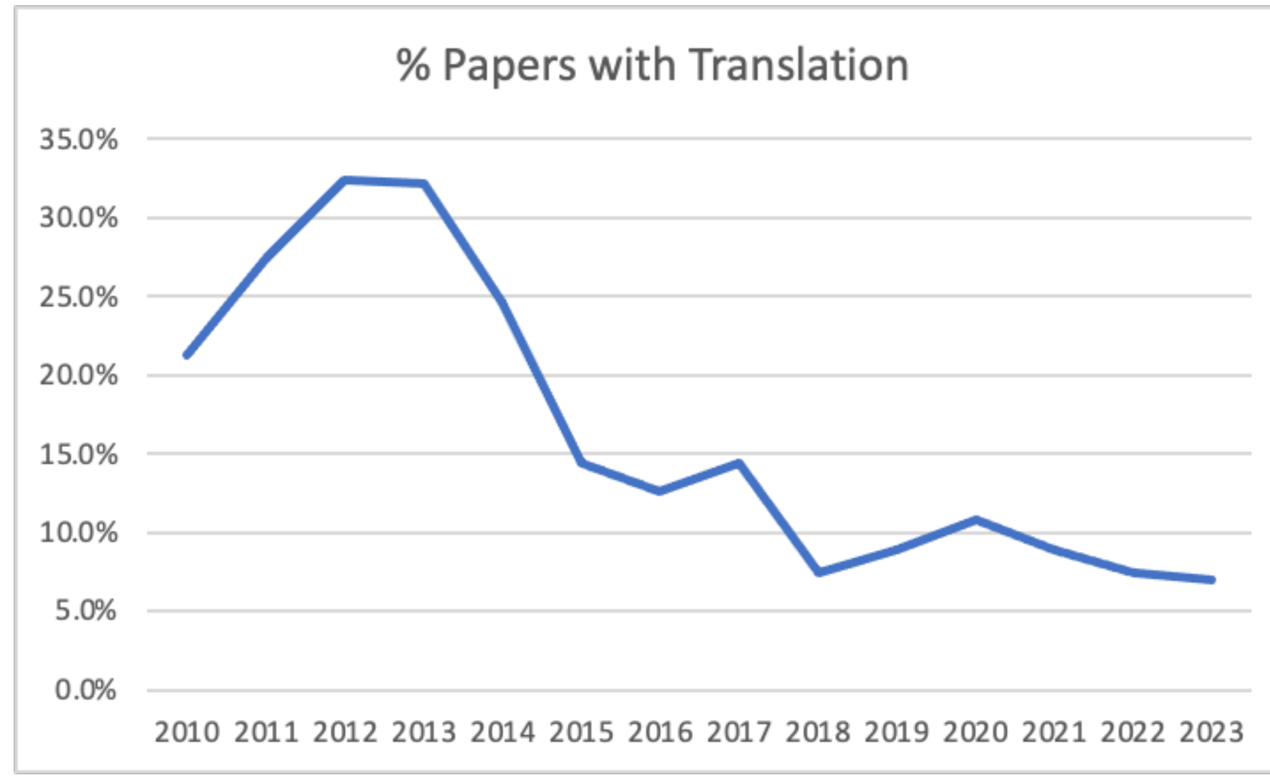
  big data, probabilistic modelling,

  encoders-decoders, attention, subwords

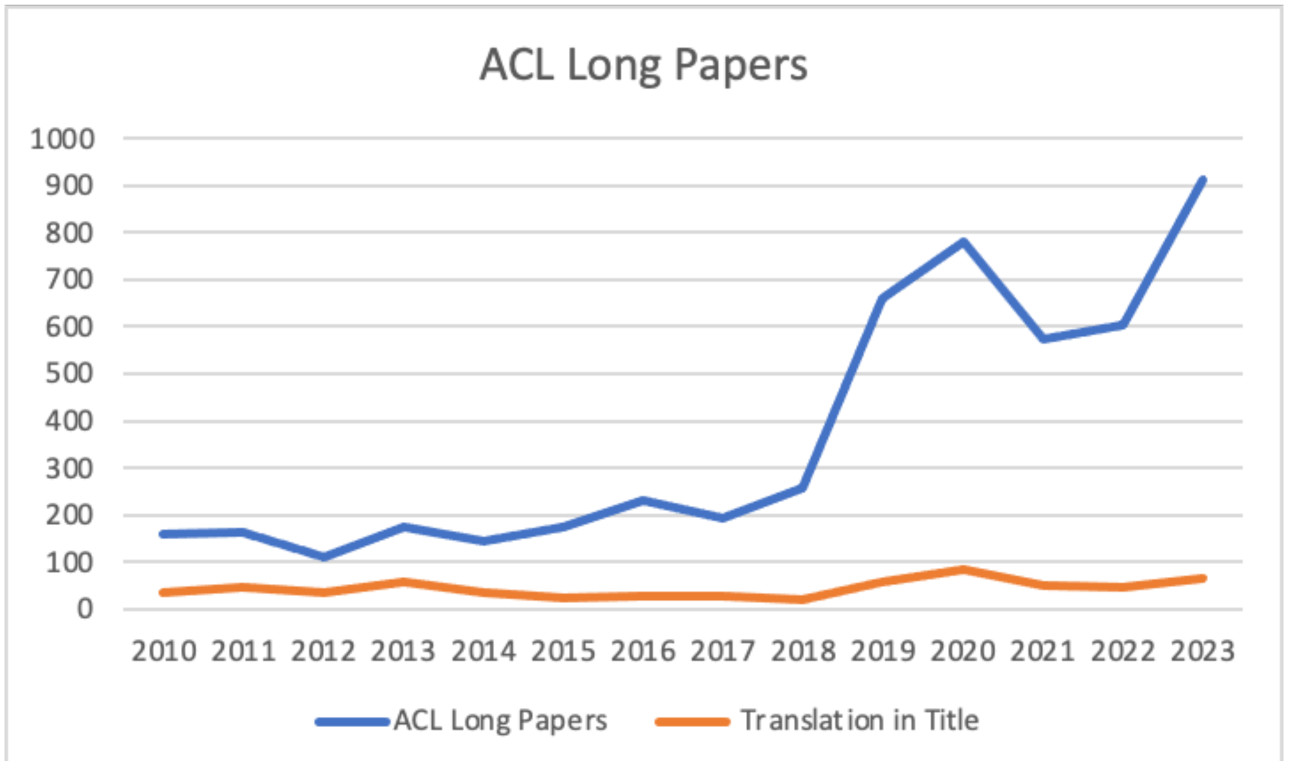
- Convergence of NLP on a unified deep learning framework - still train MT models
- And now just ask GPT: Translate "X" to Y

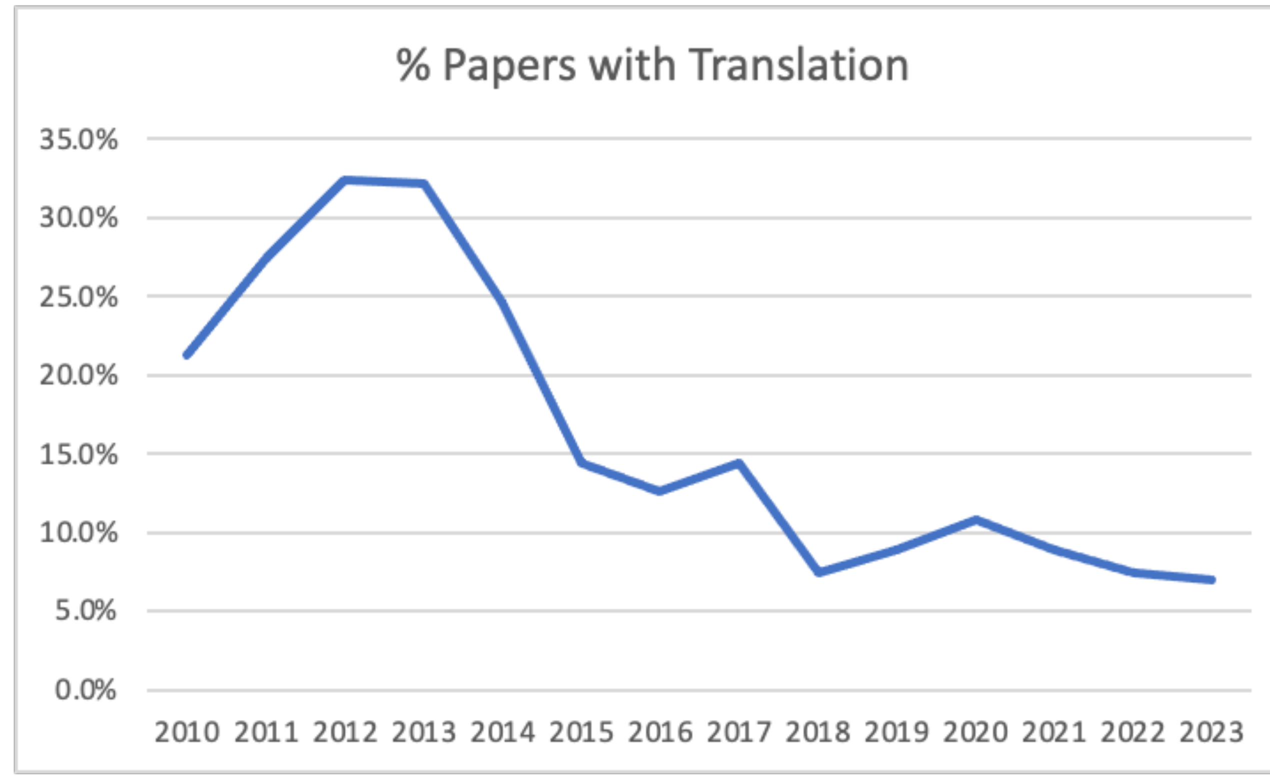
Do we need MT?


ACL Long Papers
1000
900
800
700
600
500
400
300
200
100
0
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023
ACL Long Papers
Translation in Title


% Papers with Translation
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

Do we need MT?





2010: 34

2023: 64

Neural MT

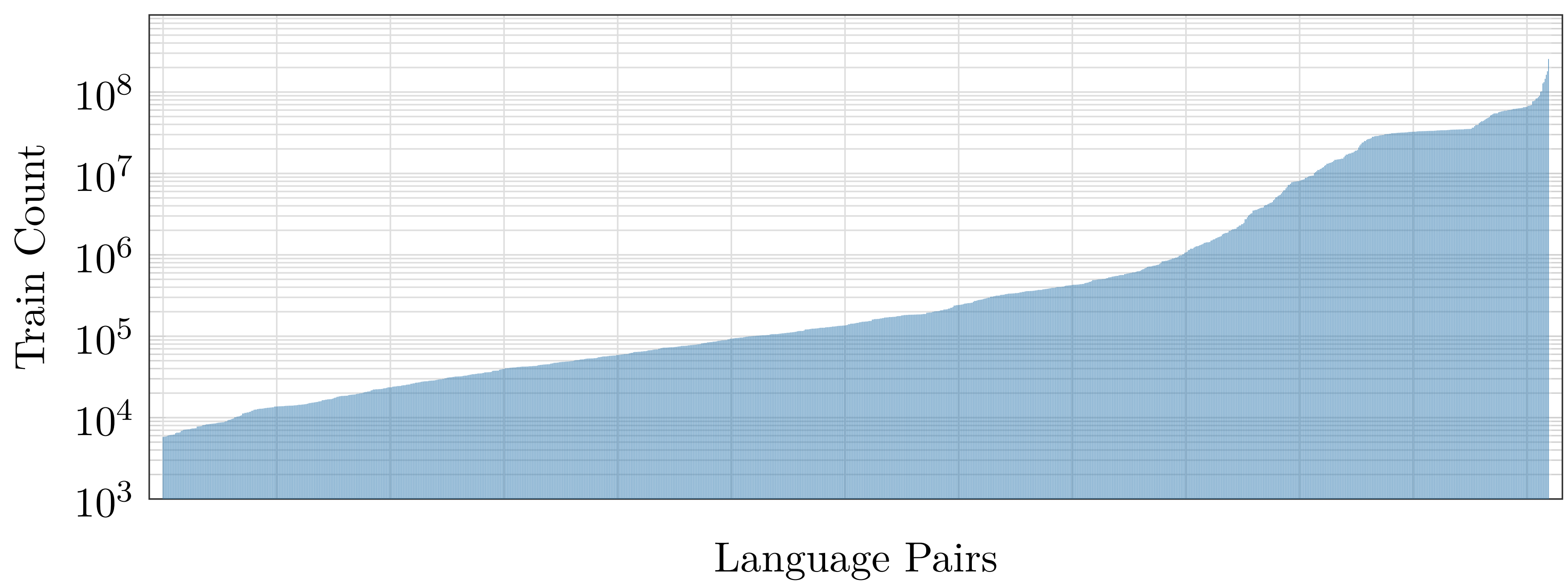
What did we mean by NMT?

- Transformer Encoder-Decoder
- Focus on parallel data
- Bilingual or Multilingual
- Large (but not that large?)
  - MBART 600M
  - 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

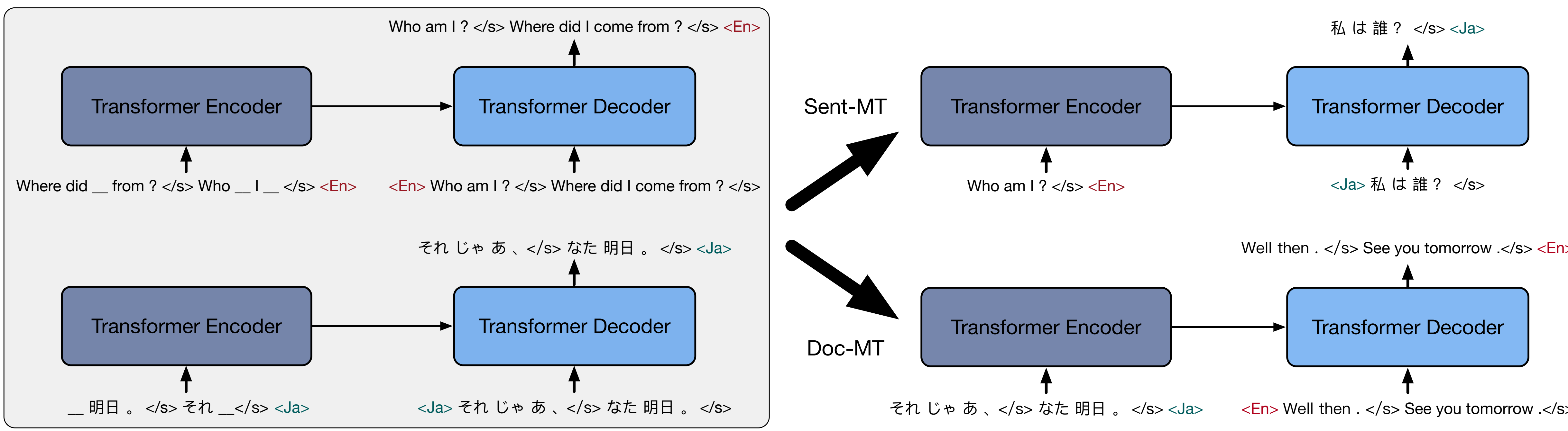


202 language, parallel/mined/BT 1220 language pairs, 18B sentence pairs

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



Multilingual Denoising Pre-Training (mBART)

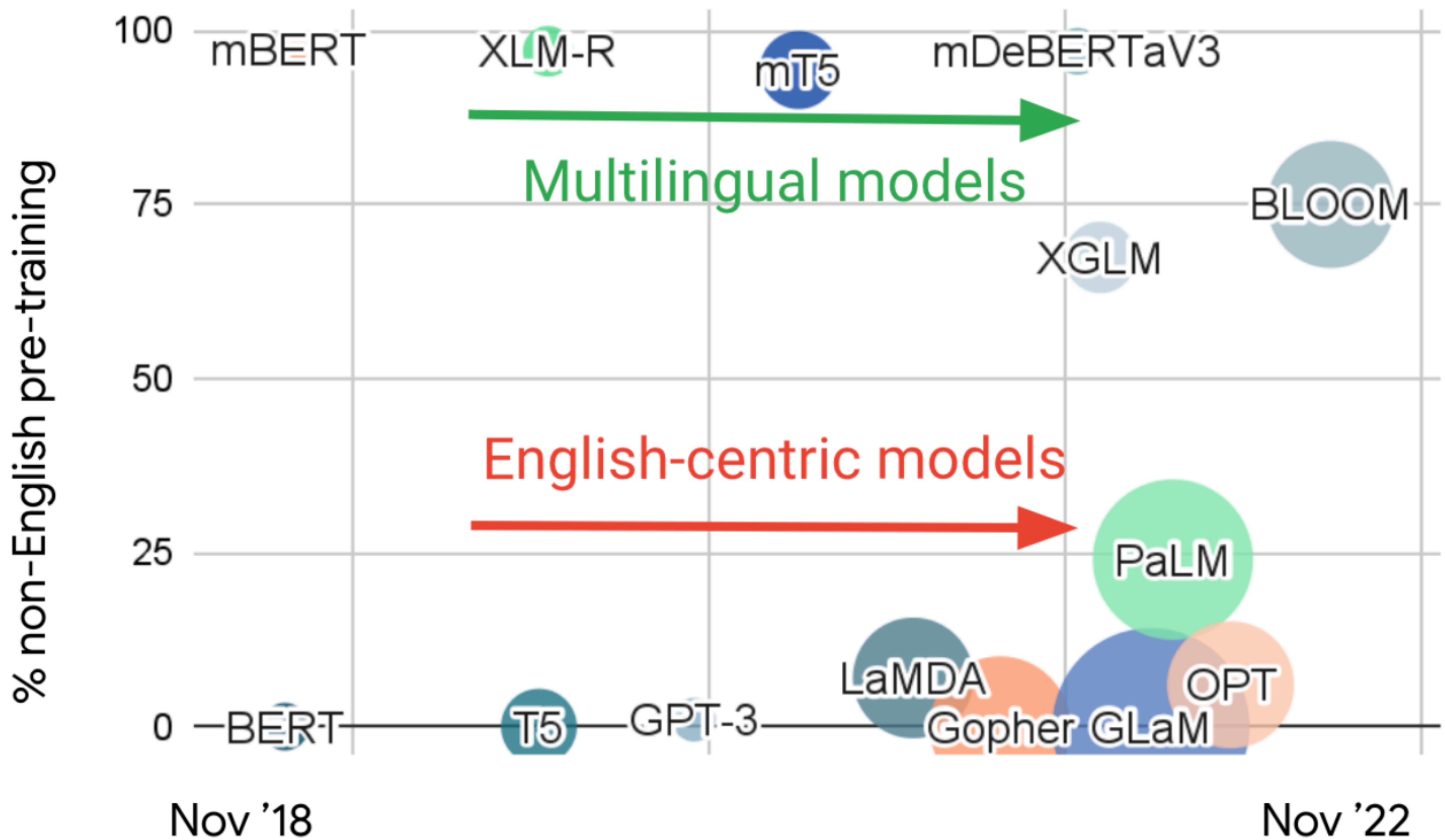
Fine-tuning on Machine Translation

600M param

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
- Mostly trained to predict next word
- Models are very large: > 7B parameters, up to 200B
- Data and compute very large - no longer in reach apart from a handful of groups

How multilingual are LLMs?



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

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]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	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	39.2	29.7	40.6	21.0	39.5

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]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	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	39.2	29.7	40.6	21.0	39.5

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]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	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	39.2	29.7	40.6	21.0	39.5

175B, 7% non English

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

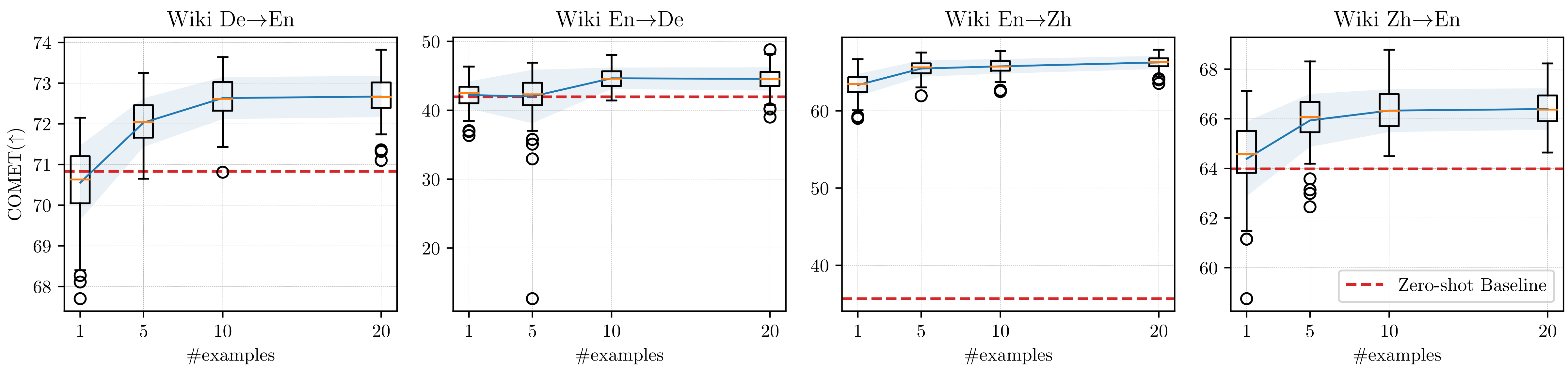
ID	Template (in English)	English		German		Chinese	
		w/o	w/	w/o	w/	w/o	w/
A	[src]: [input] ◇ [tgt]:	**38.78**	**31.17**	-26.15	-16.48	**14.82**	**-1.08**
B	[input] ◇ [tgt]:	-88.62	-85.35	-135.97	-99.65	-66.55	-85.84
C	[input] ◇ Translate to [tgt]:	-87.63	-68.75	-106.30	-73.23	-63.38	-70.91
D	[input] ◇ 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
1-Shot Translation (high-quality pool)				
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
5-Shot Translation (high-quality pool)				
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
1-shot Translation (Low-quality Pool)				
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
1-Shot Translation (high-quality pool)				
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
5-Shot Translation (high-quality pool)				
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
1-shot Translation (Low-quality Pool)				
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
1-Shot Translation (high-quality pool)				
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
5-Shot Translation (high-quality pool)				
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
1-shot Translation (Low-quality Pool)				
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 BLEU	Wiki COMET	WMT BLEU	WMT COMET
Zero-Shot	24.08	33.92	20.38	17.97
1-Shot Translation (high-quality pool)				
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
5-Shot Translation (high-quality pool)				
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
1-shot Translation (Low-quality Pool)				
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
1-Shot Translation (high-quality pool)				
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
5-Shot Translation (high-quality pool)				
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
1-shot Translation (Low-quality Pool)				
Random	24.75	38.86	22.06	30.70
Ours	24.94	39.88	22.23	30.87

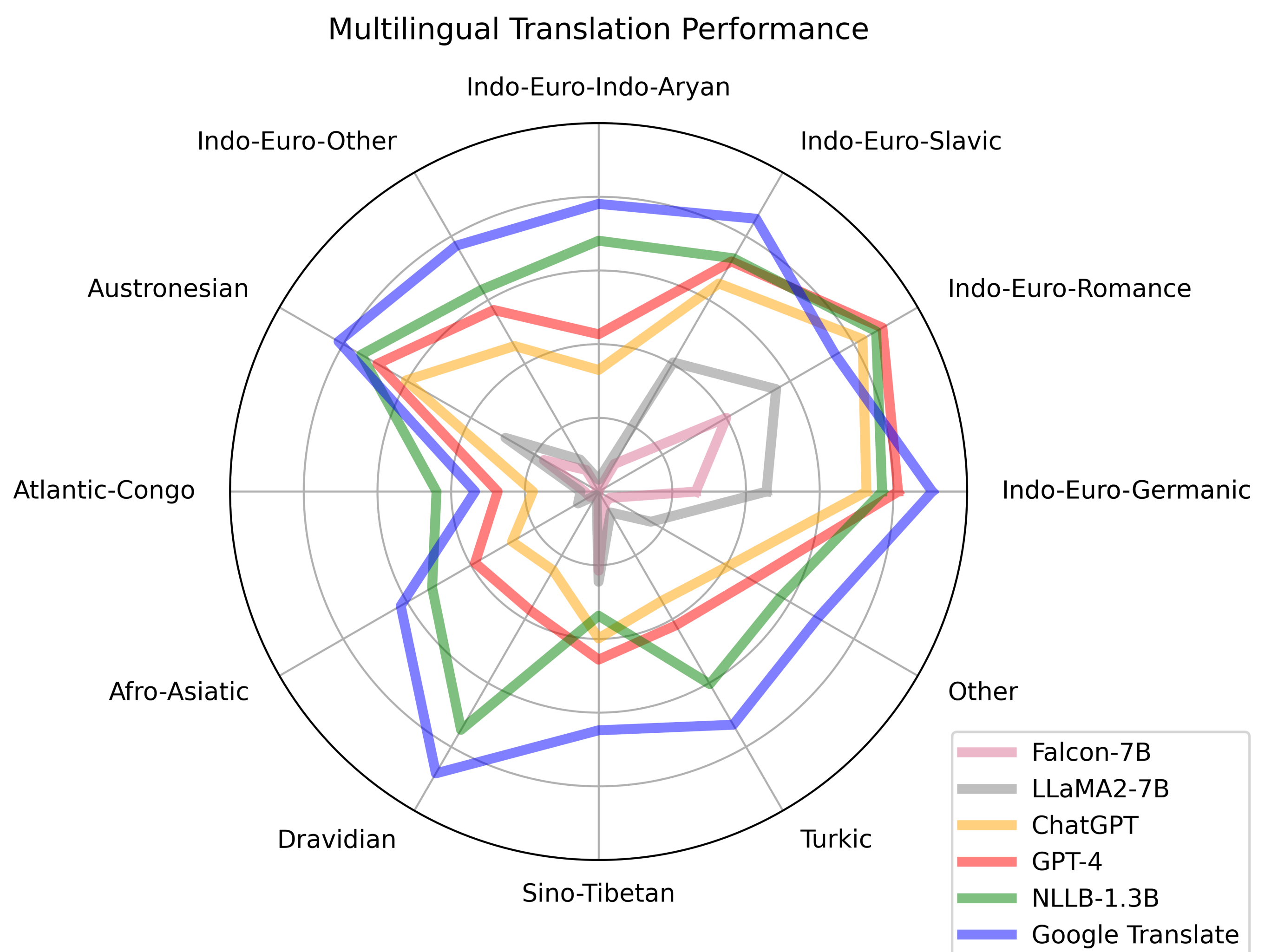
Problems remain

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 三江源国家公园管理局长江源园区可可西里管理处, the total number of re-migration of the Tibetan antelope has been on the rise since 2018, with 4,860 re-migrating in 2109, an increase of 338 compared to 2808.
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**

Are LLMs competitive?

Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis
Wenhao Zhu et al. 2023



En-X, 8 in context examples, 101 Flores languages

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

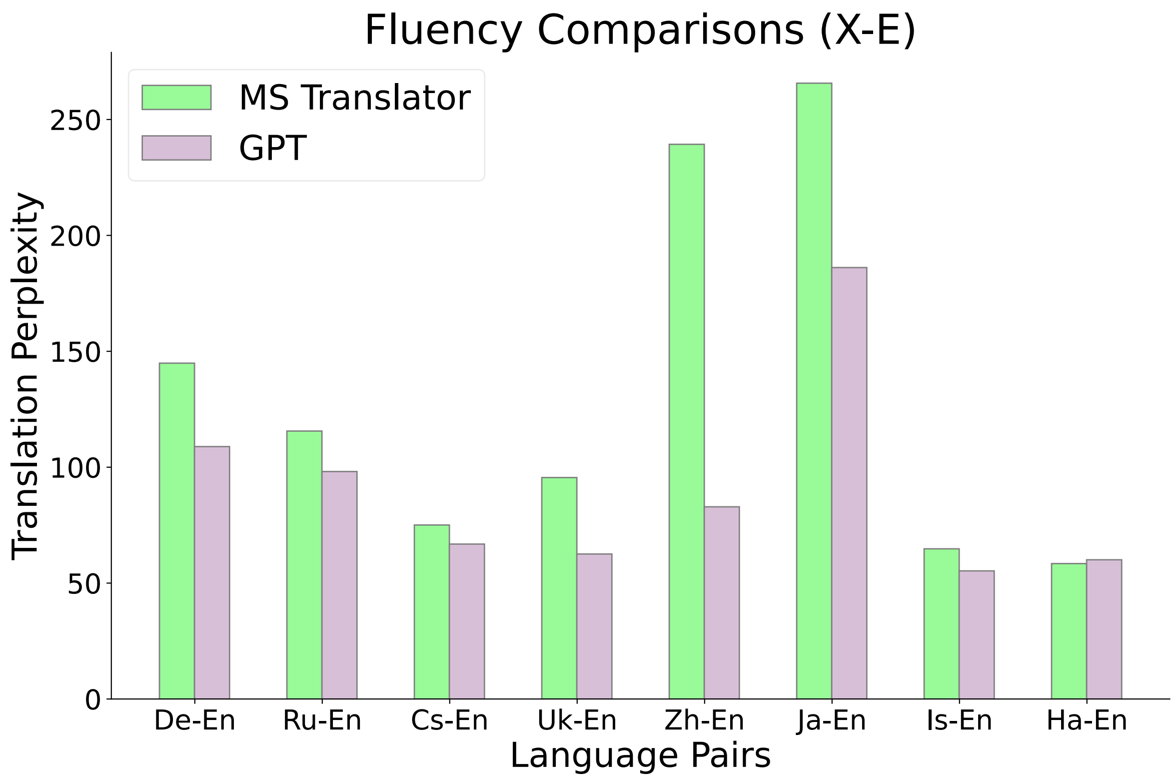
Monolingual Bias

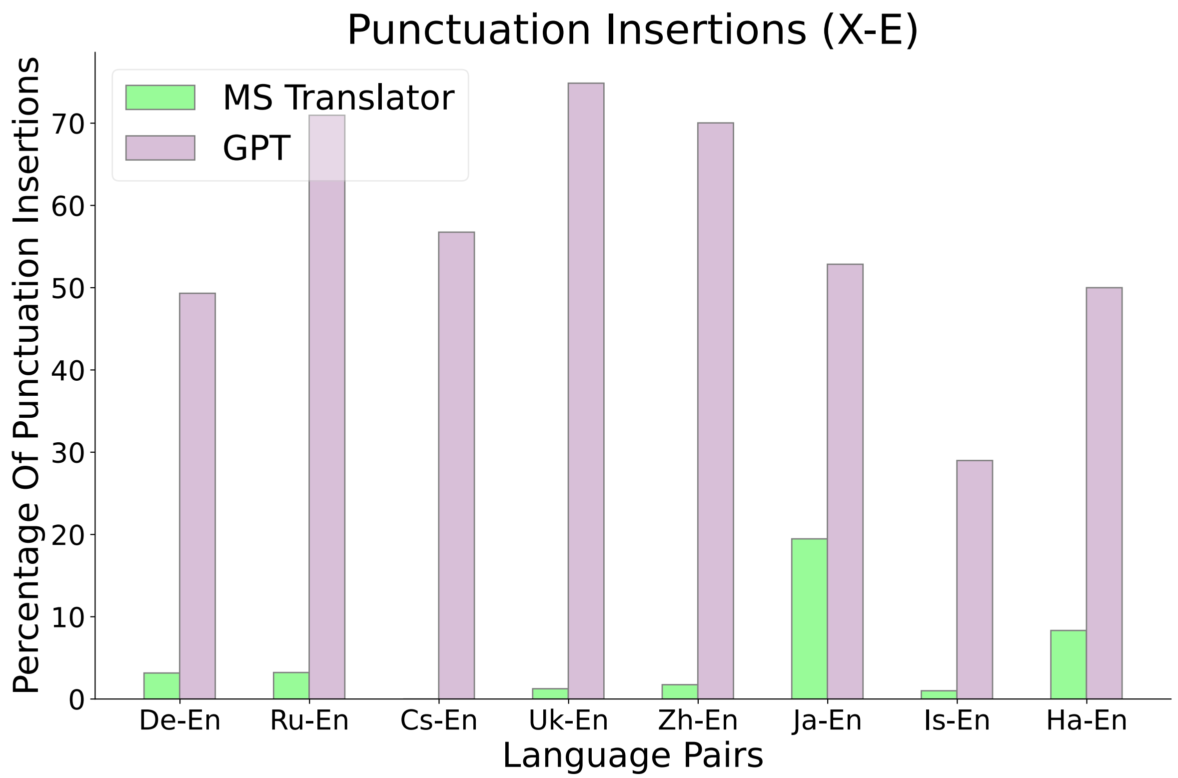
- Instructions might fail to override LLM training
- Lack of teacher forcing supervision means might not be faithful to source sentence
- Favour fluency over accuracy eg, introducing undesirable punctuation or removing tokens which have been unseen

Behaviour LLMs vs MT?

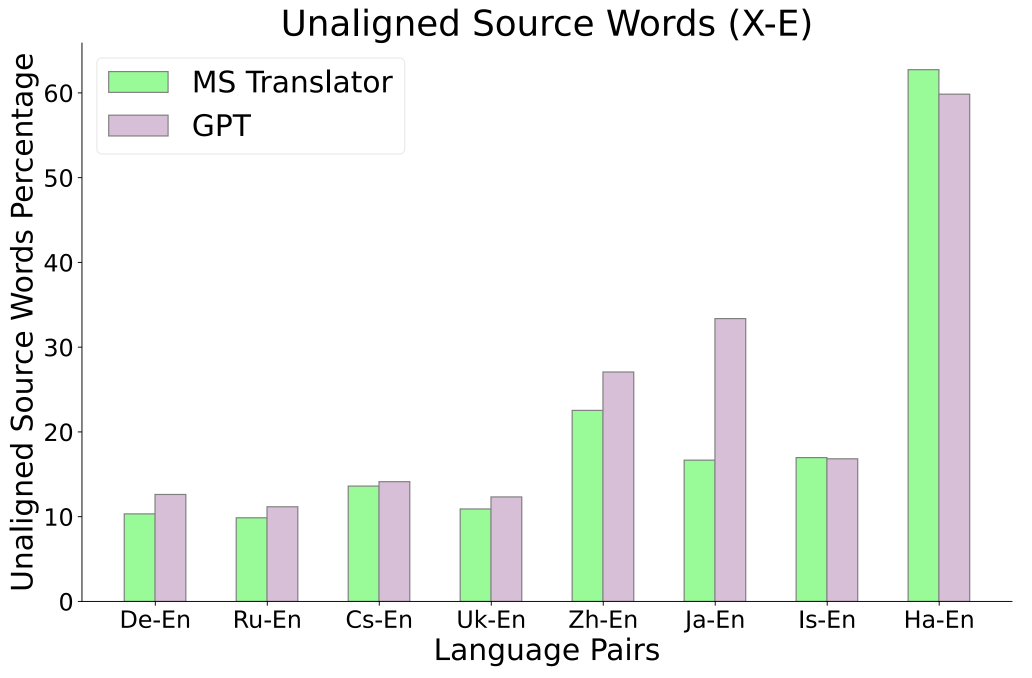
Sequence Type	Translation Instance	Phenomenon
Source MS Translator GPT	Bis auf die E 95 02 wurden **alle Lokomotiven zerlegt**. With the exception of E 95 02, **all locomotives were dismantled**. **All locomotives were dismantled** except for the E 95 02.	**Non-Monotonicity (NM)**
Source MS Translator 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)**

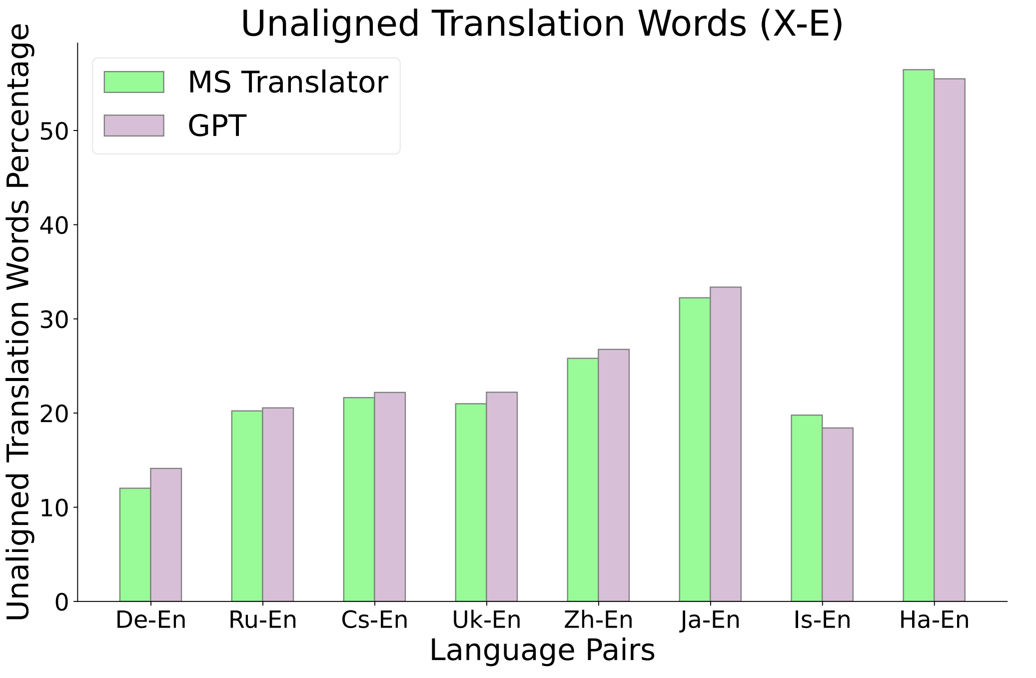
Behaviour LLMs vs MT?

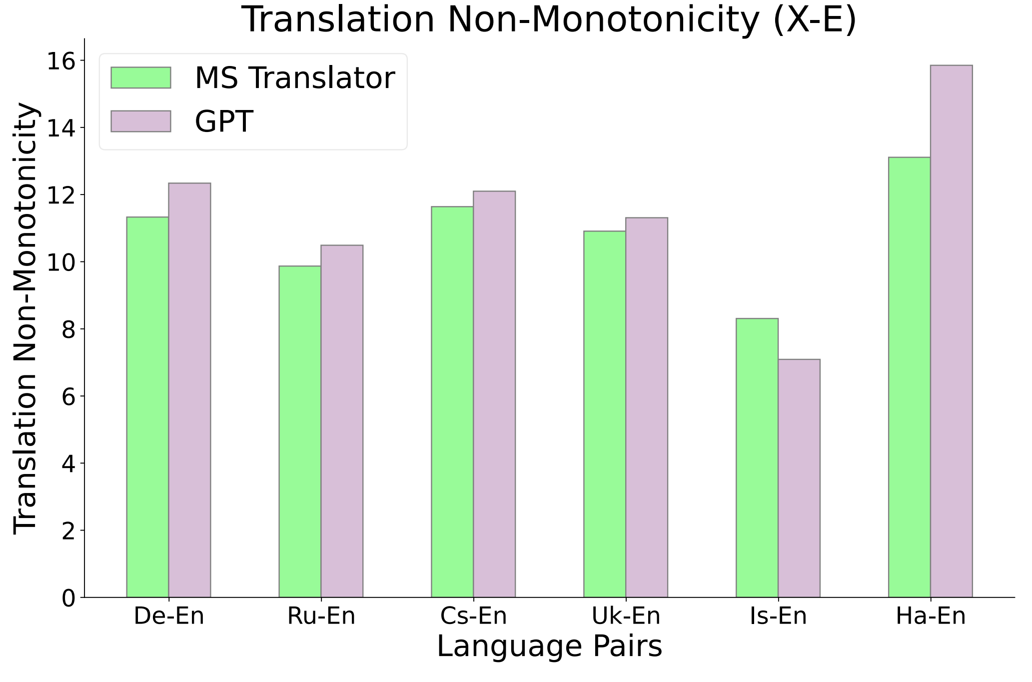




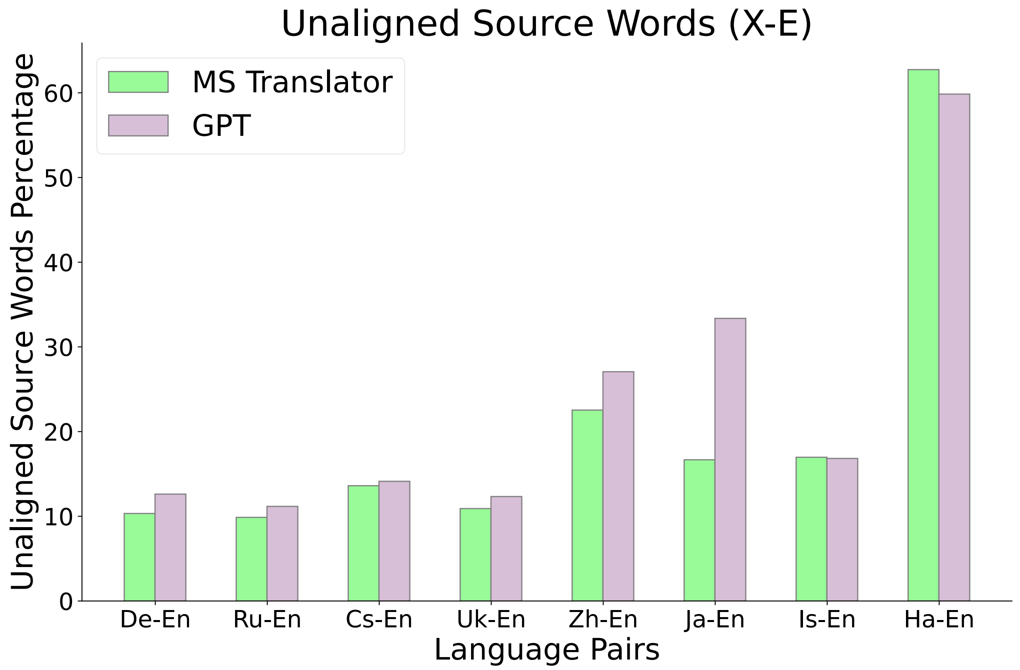
Behaviour LLMs vs MT?

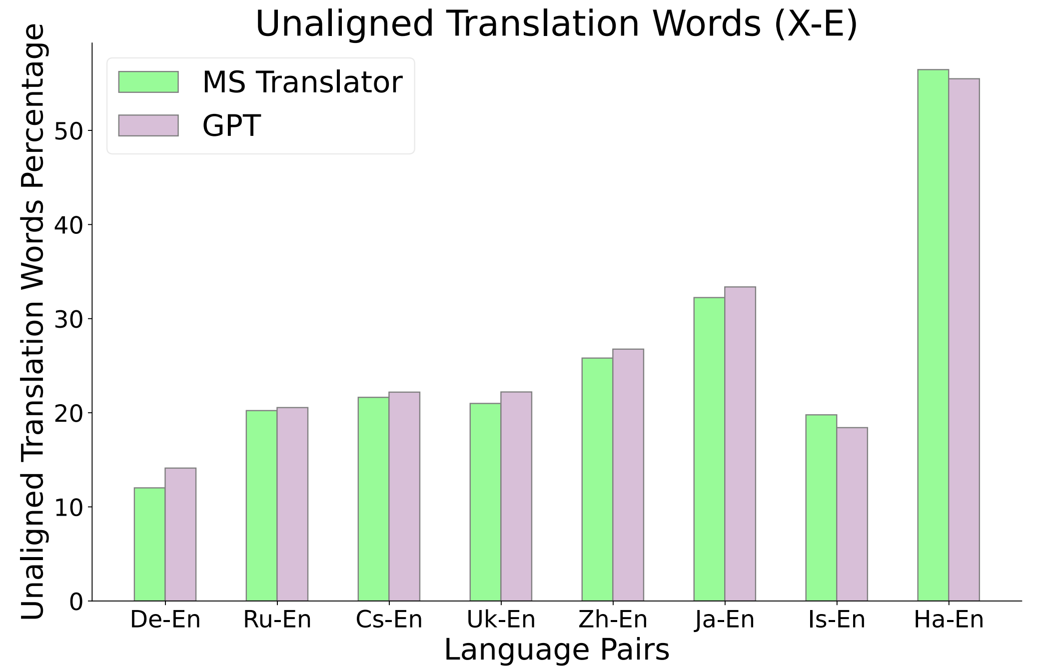


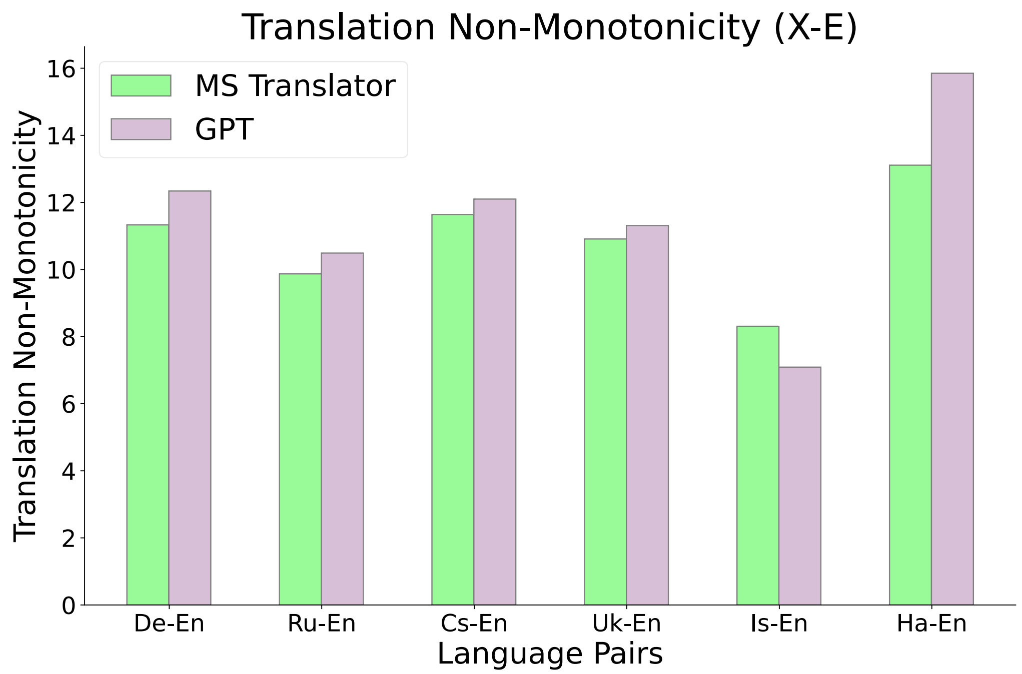




Behaviour LLMs vs MT?





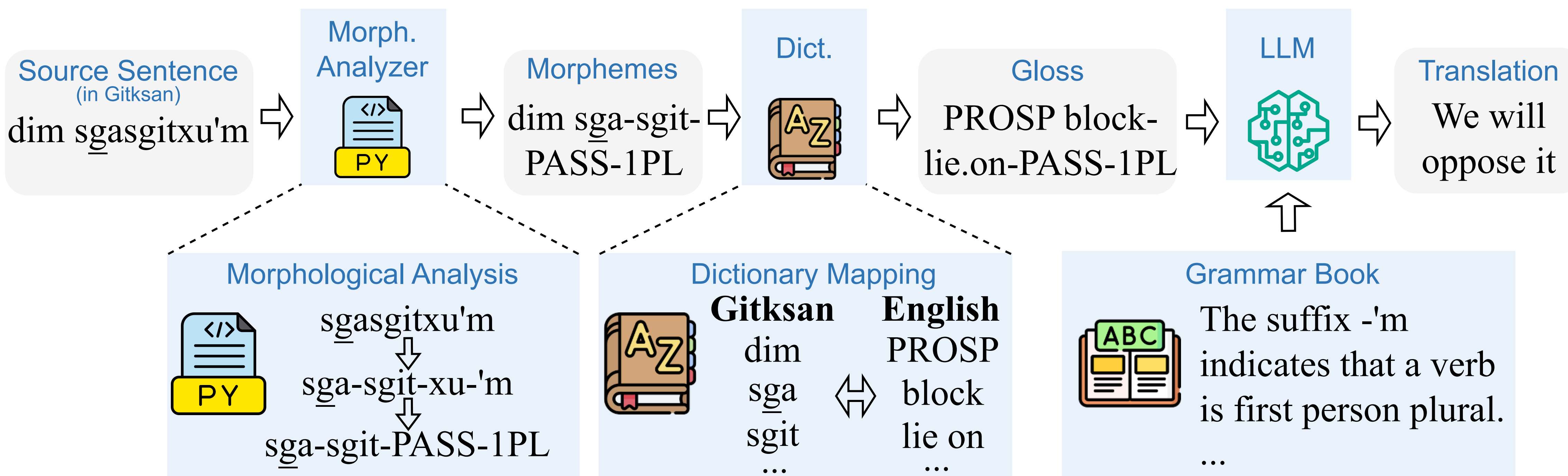


- Ignore more source - no more inserted target
- Less literal
- Better for figurative text

Opportunities

Controllable:

Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions. Zhang et al. 2024



Opportunities

Controllable:

Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions. Zhang et al. 2024

	mnc →en	git →en	usp →es	ntu →en	ddo →en	wol →en	arp →en	arp en→	bzd →es	bzd es→	Avg.
GPT-4											
Zero-Shot	0	0	0.1	0	0	3.9	0	0.2	0.4	0	0.5
Zero-Shot CoT	0.7	0	0.3	0	0	11.4	0.4	4.1	0.4	0.1	2.4
Few-Shot	0.5	9.3	2.2	0	0.8	**13.5**	1.0	2.2	0.8	1.7	3.2
LingoLLM dict. only	8.3	7.7	10.7	11.7	11.1	6.9	6.0	14.5	2.7	2.2	8.2
LingoLLM	**10.8**	**14.3**	**12.4**	**12.9**	**15.1**	8.1	**9.4**	**15.6**	**4.3**	**3.0**	**10.5**
Mixtral-8x7B											
Zero-Shot	0.2	2.0	0.3	1.2	0.8	7.4	0.8	0.5	0.2	0	1.3
Zero-Shot CoT	0.5	3.4	0.2	1.3	0.4	6.2	0	0.7	0.5	0.1	1.3
Few-Shot	0.5	4.0	2.2	2.2	0.6	8.6	0.9	0.5	1.7	1.8	2.3
LingoLLM dict. only	4.1	4.7	3.9	6.3	6.0	6.0	5.2	7.3	2.6	1.3	4.7
LingoLLM	**4.4**	**7.9**	**4.6**	**7.3**	**10.7**	3.2	**7.4**	**8.4**	**3.0**	**2.2**	**5.9**

Opportunities

Controllable:

Towards Effective Disambiguation for Machine Translation with Large Language Models. Iyer, Chen and Birch 2023

Source	The horse had a blaze between its eyes.
DeepL	那匹马的两眼之间有一团火焰。 (There is a flame between the horse's eyes.)
BLOOMZ (176B)	这匹马的眼睛之间有一道白线。 (There is a white line between the horse's eyes.)

Opportunities

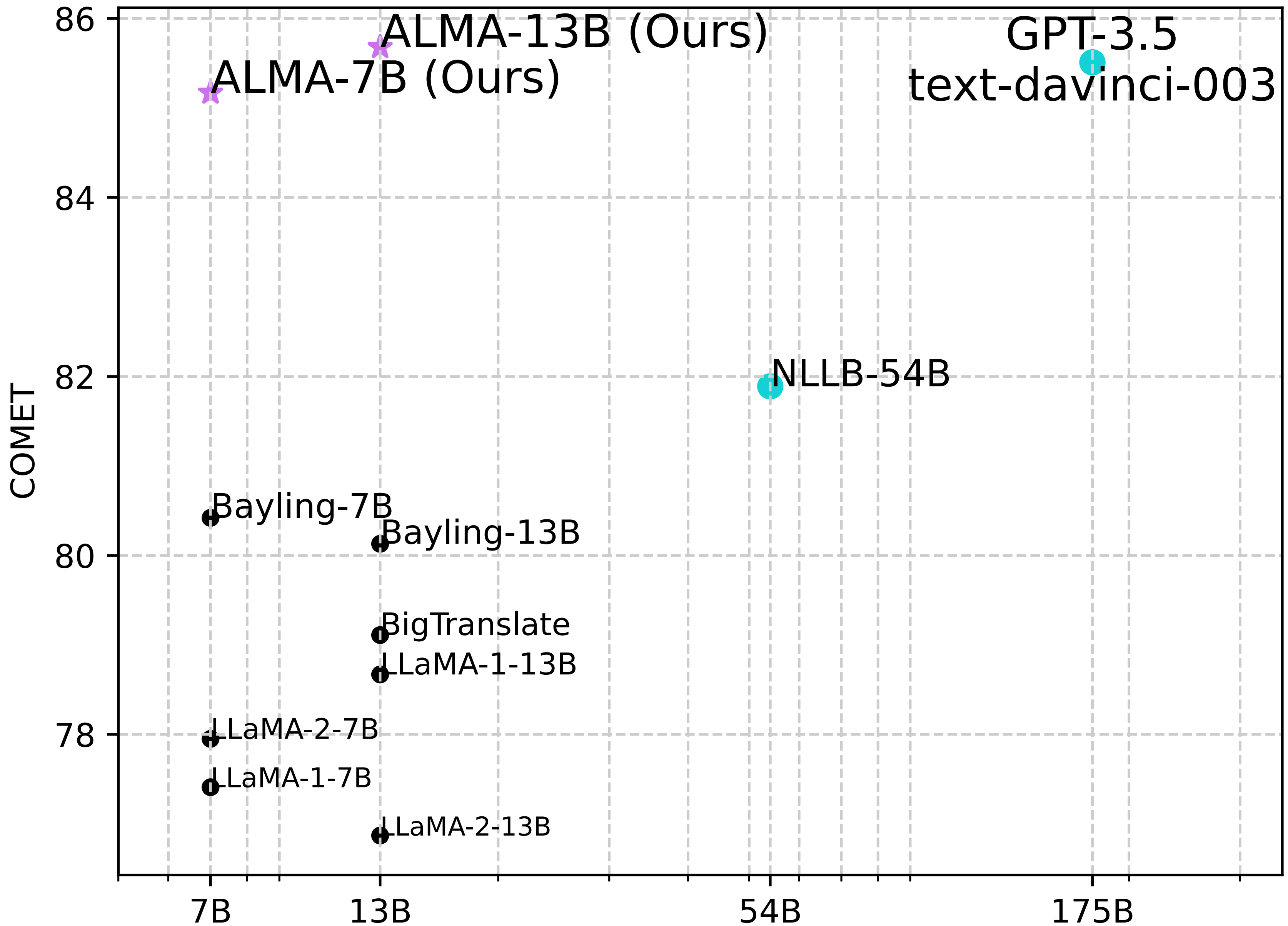
Controllable:

Towards Effective Disambiguation for Machine Translation with Large Language Models. Iyer, Chen and Birch 2023

System	1-shot		3-shot		5-shot	
	Rand.	Sim.	Rand.	Sim.	Rand.	Sim.
DeepL			—63.91—			
NLLB-200 54B			—***61.33***—			
LLaMA 7B	53.64	**54.01**	**55.53**	52.52	**56.33**	54.45
LLaMA 65B	56.57	**59.38**	59.83	**62.44**	60.78	**63.74**
BLOOM 176B	**63.66**	62.44	64.52	**66.19**	65.53	**68.22**
BLOOMZ 176B	64.35	**69.57**	67.31	**71.15**	68.55	**71.33**

Pretrain-FineTune-Prompt Paradigm

A Paradigm shift in machine translation: Boosting translation performance of large language models. Haoran Xu et al. 2023



So do we need MT?

- MT models and LLM models have somewhat converged with differences in: size, encode/decoder, amount of monolingual and parallel data in pre-training and finetuning

- LLMs are robust, controllable and produce excellent MT performance when fine tuned, need far less parallel data

- But MT models use far less data overall, and are much smaller, use less compute to train and run and produce highest quality literal translations in the right language

Conclusion

- Saying: "May you live in interesting times"
- A huge number of research problems now possible on the border between translations and generation
- MT 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
  - Compelling task that can benefit humanity