Tingchen Fu is currently looking for Ph.D position staring at 25fall.

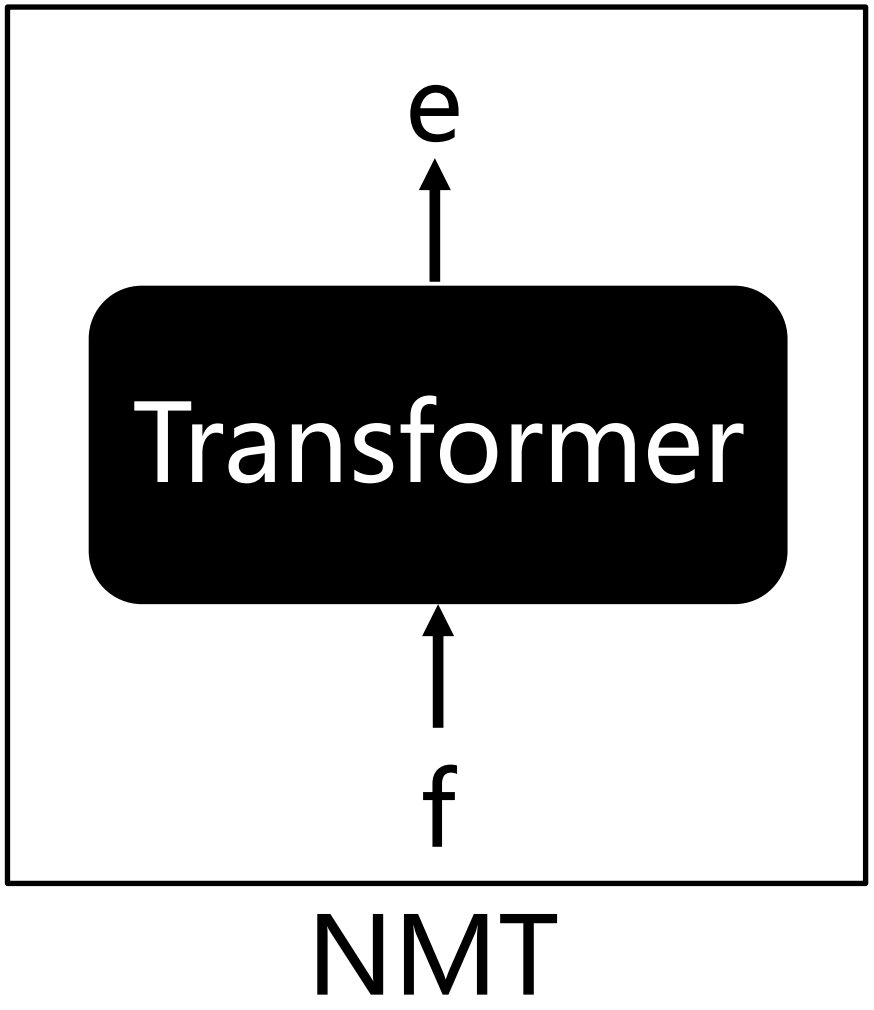
Tencent AI Lab

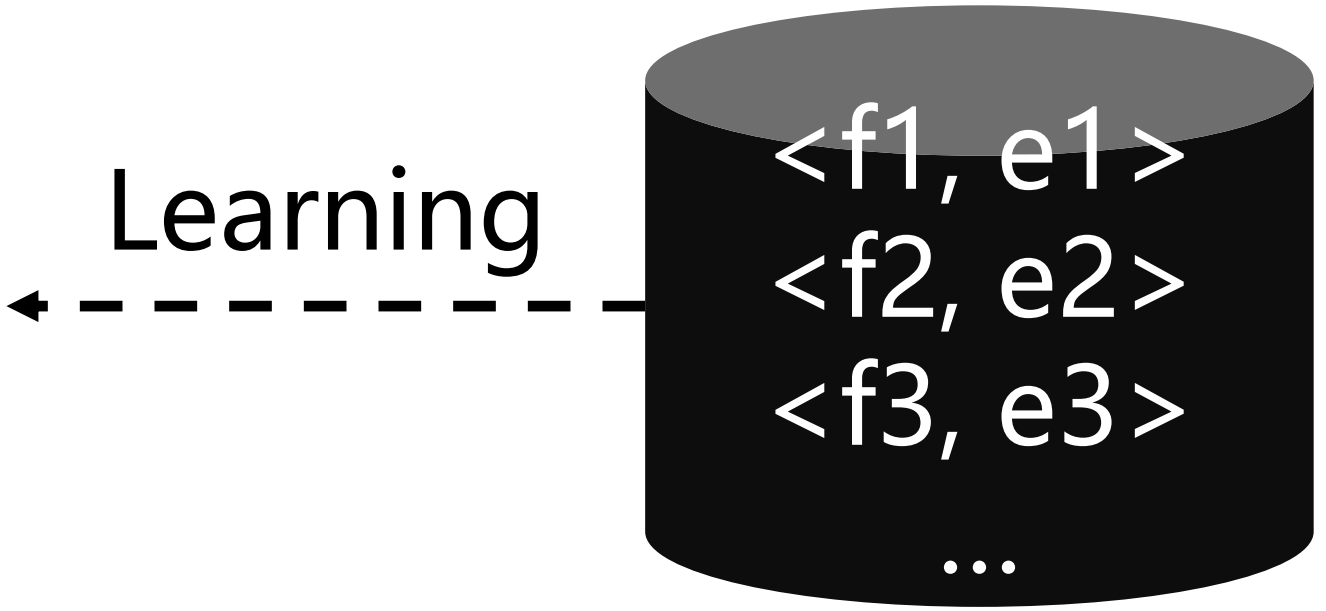
The Reasonableness Behind Unreasonable Translation Capability Of Large Language Model

Tingchen Fu
Renmin University of China
Intern@Tencent AI Lab

ICLR 2024: https://openreview.net/pdf?id=3KDbIWT26J

Translation capability from bilingual data

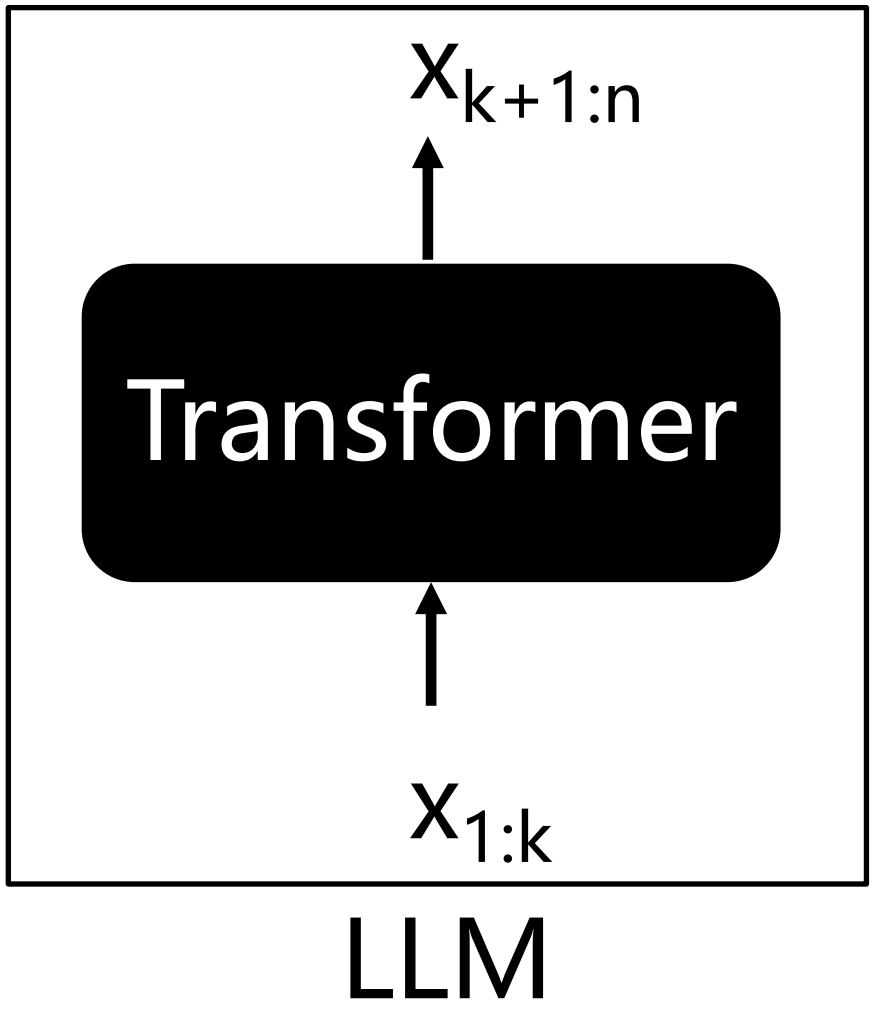


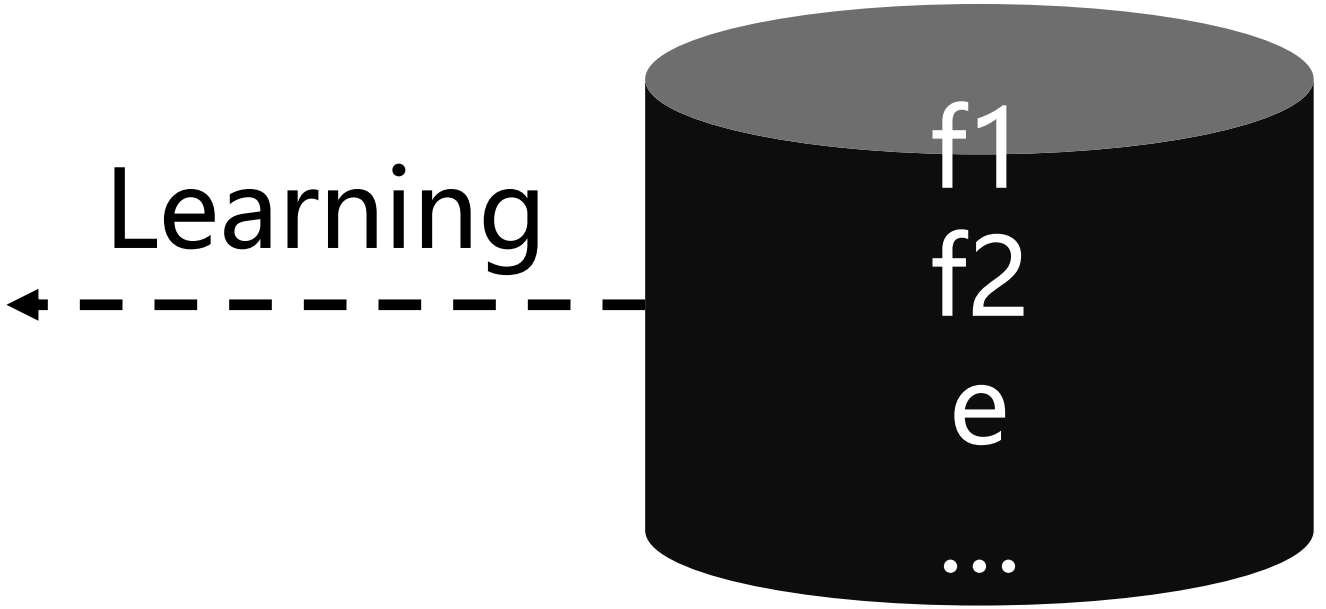


	Chinese⟶English		English⟶German	
	BLEURT ↑	MQM (Human) ↓	BLEURT ↑	MQM (Human) ↓
Google Translate	68.5	3.1	73.0	1.0

[Google 23] PaLM 2 Technical Report.

Translation capability from monolingual data





	Chinese⟶English		English⟶German	
	BLEURT ↑	MQM (Human) ↓	BLEURT ↑	MQM (Human) ↓
PaLM	67.4	3.7	71.7	1.2
Google Translate	68.5	3.1	73.0	1.0
PaLM 2	**69.2**	**3.0**	**73.3**	**0.9**

[Google 23] PaLM 2 Technical Report.

[Jiao et al. 23] Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine

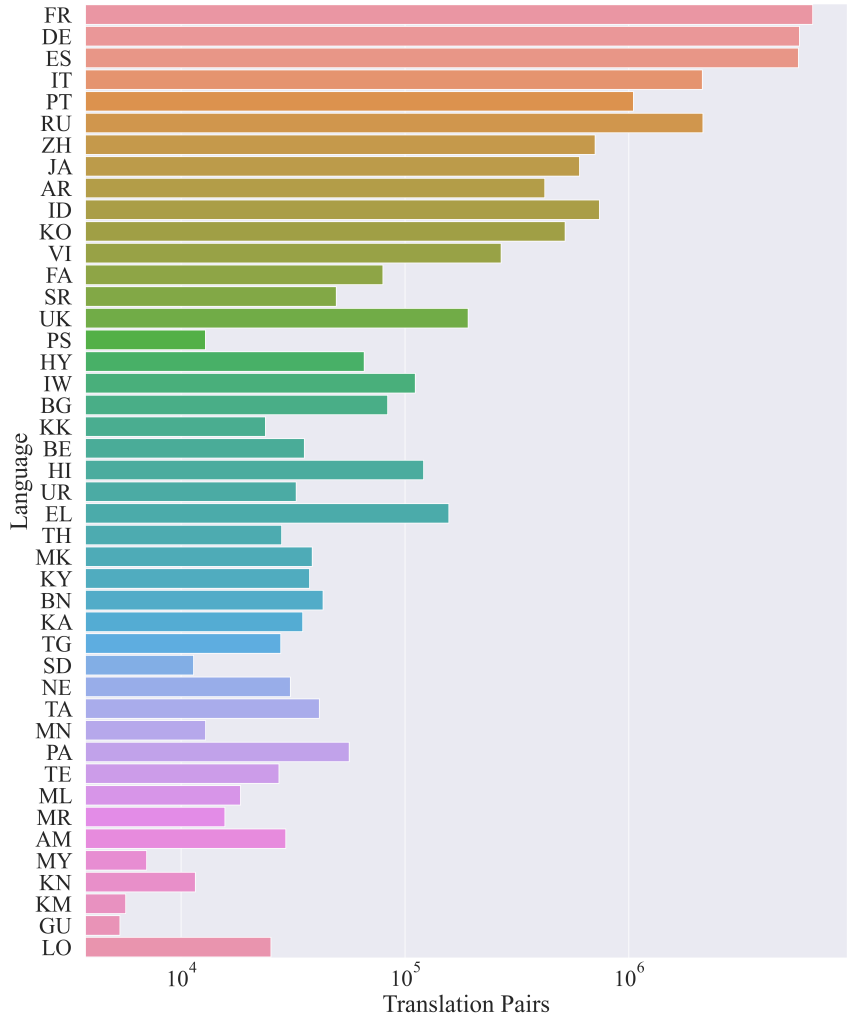
Why LLMs enable Translation capability?

- Monolingual data contains some parallel data

Example
This news, like a light as an indescribable speed, In the blink of an eye it spread throughout the entire Martial Dragon Continent. *这个消息，如同光芒一般，以无法形容的速度，眨眼间就传遍了整个龙武大陆。*

- The included parallel data is able to train a strong NMT model.

t	#TRANSLATIONS	PaLM (mined)	WMT
N/A	40,836,876	✗	42.0
0.90	9,084,429	33.7	
0.80	7,056,441	35.7	
0.70	4,874,173	36.4	
0.60	3,341,187	37.3	38.1
0.50	2,474,703	37.2	
0.40	1,948,820	37.1	
0.30	1,477,535	38.4	36.5
0.20	906,937	37.8	
0.15	549,705	36.3	



[Google ACL 23] Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM’ s Translation Capability.

Why LLMs enable Translation capability?

Tencent AI Lab

- However, LLM still yields strong translation capability without parallel data.

BLEU score of PaLM through 5-shot learning

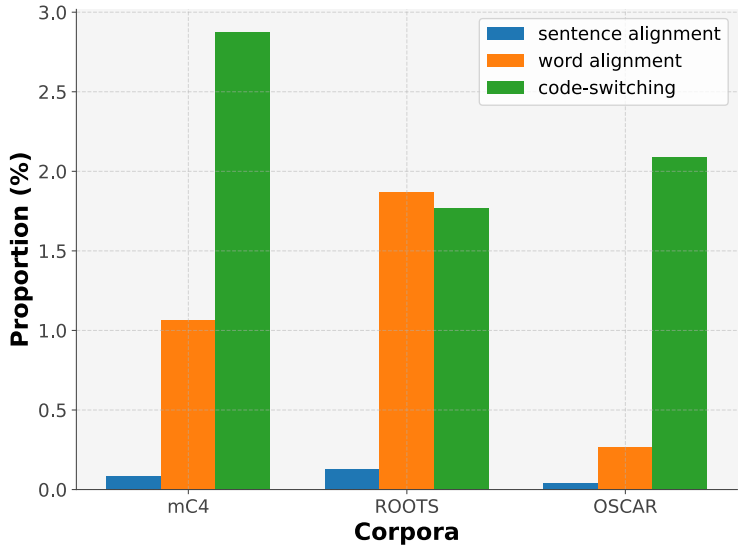
Models	EN-XX		XX-EN	
	FULL	-PAR	FULL	-PAR
1B	30.9	18.7	12.5	5.1
7B	47.7	44.7	24.0	22.2

- Why LLM capture translation capability without parallel data?
  - This is the focus of our work!

[Google ACL 23] Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM' s Translation Capability.

Possible factors on translation capability of LLM

- Some data sources may be related to translation capability
  - SA: sentence alignment
  - WA: word alignment
  - CS: code-switching

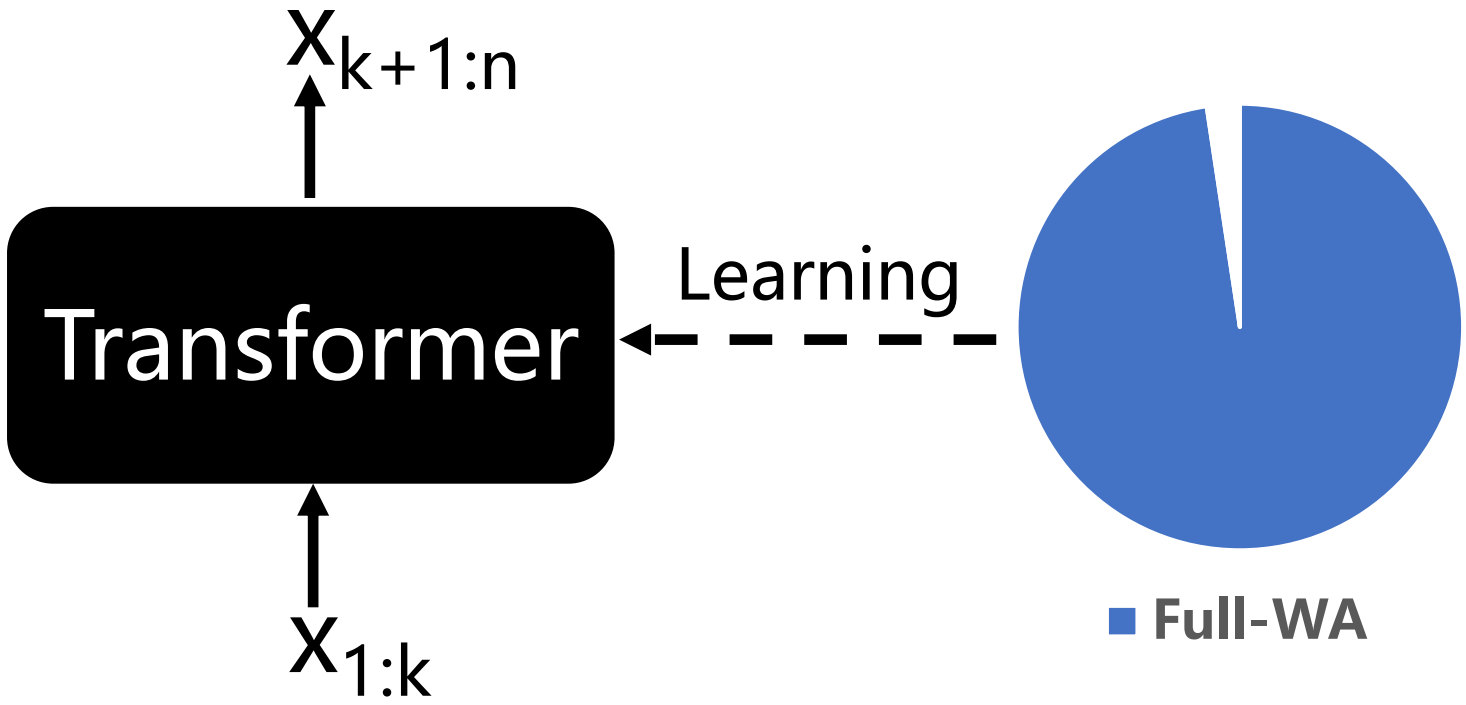


Type	Example
Sentence Alignment	*This news, like a light as an indescribable speed, In the blink of an eye it spread throughout the entire Martial Dragon Continent.* *这个消息，如同光芒一般，以无法形容的速度，眨眼间就传遍了整个龙武大陆。* This news was like a bullet, landed on the tranquil lake in the middle, instantly exploded!
Word Alignment	Beijing will procure RMB 80 million in social organization services. *Beijing News* (*新京报*), January 28, 2013
Code-Switching	上一篇 ： New Polio Immunization Drive to Start in Nigeria’s 下一篇 ： Hong Kong's Top Health Official Resigns Over SARS

Naïve Method and Challenges

Tencent AI Lab

Re-training a 1b or 7b model on the corpus excluding particular data (e.g. word alignment) ?



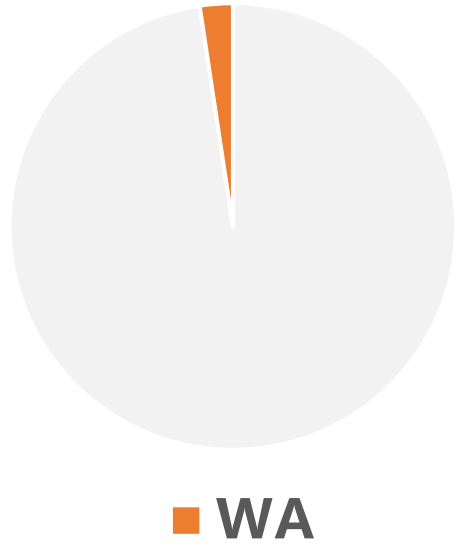
Models	EN-XX	
	FULL	Full-WA
1B	30.9	xx
7B	47.7	xx

- Challenge: it is too expensive!
  - Training a BLOOM-1b requires several months on 300+ A100 GPUs.
- Efficient methods?

Efficient Methods

Tencent AI Lab

- Finetuning an off-the-shelf LLM model
  - Finetuning the model on WA (or CS)
  - Finetuning the model on random data (with the same size)
  - Evaluating both models in terms of translation quality
- Training small models from scratch as simulation
  - Training 560m model on WA (or CS)
  - Training 560m model on random data
  - Evaluating both models in terms of translation quality



Experiments — Data Preparation

Tencent AI Lab

- Training data from mC4 dataset
  - SA: sentence alignment
  - WA: word alignment
  - CS: code-switching
- Evaluation datasets
  - WMT21
  - FLORES-200

		mC4.en	mC4.zh
sentence alignment	# Doc	210,931	2,462
	# Seq	355,320	432
word alignment	# Doc	658,643	1,972,764
	# Seq	500,550	659,456
code-switching	# Doc	2,021,502	5,086,373
	# Seq	903,810	997,376

Dataset	Language	Test set	Example pool
WMT21	English-Chinese	newstest2021 (1948/1002)	newstest{2017,2018,2019}
FLORES-200	English	eng_Latn.devtest (1012)	eng_Latn.dev (997)
	Chinese	zho_Hans.devtest (1012)	zho_Hans.dev (997)
	Catalan	cat_Latn.devtest (1012)	cat_Latn.dev (997)
	Eastern Panjabi	pan_Guru.devtest (1012)	pan_Guru.dev (997)
	Igbo	ibo_Latn.devtest (1012)	ibo_Latn.dev (997)
	Tswana	tsn_Latn.devtest (1012)	tsn_Latn.dev (997)

Experiments — Settings

Tencent AI Lab

- Finetuning
  - Using the bloom models 7b and 560m as initialization
  - Finetuning on SA/WA/CS with one epoch for fair comparison
- Training a small model for simulation
  - Training a 560m model from scratch
  - Training the model with a fixed number of updates

Experiments — Finetuning a 7b model

Tencent AI Lab

X-random data denotes the same number of examples as X dataset randomly sampled from the training data of LLM for X in {SA, WA, CS}

	ZH-EN				EN-ZH			
	3-shot		5-shot		3-shot		5-shot	
	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT
BLOOM-7.1b	59.58	37.21	60.38	38.01	79.84	57.87	80.34	58.58
SA	62.05*	41.24*	61.79*	40.47*	79.77	58.32*	80.18	58.64
SA-rand	59.13	37.60	59.28	37.73	79.47	57.48	79.99	58.33
WA	58.36*	36.34*	58.15*	35.75*	79.59	57.61	80.11*	58.46
WA-rand	56.21	32.91	56.51	33.32	79.48	57.42	79.86	58.14
CS	60.00*	38.59*	59.54*	37.82*	78.59	56.63	79.48	57.53
CS-rand	56.64	33.39	57.50	34.44	79.20	57.34	80.24	58.30

- WA and CS may provide translation signals to LLMs
  - WA or CS contains more translation signals than random data
- WA is worse than BLOOM-7.1b
  - WA does not contain more translation signals than BLOOM-7.1b

Experiments — Finetuning a 560m model

Tencent AI Lab

	ZH-EN				EN-ZH			
	3-shot		5-shot		3-shot		5-shot	
	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT
BLOOM-560m	53.62	34.00	54.55	35.14	66.84	43.23	67.88	44.40
SA	61.55*	43.14*	61.57*	43.04*	69.27*	46.32*	69.98*	47.14*
SA-rand	54.87	36.41	55.26	36.60	61.80	38.58	63.71	40.33
WA	60.99*	42.09*	60.77*	41.72*	71.82*	49.03*	72.47*	50.24*
WA-rand	58.47	37.83	57.47	36.39	67.55	43.44	68.23	44.27
CS	59.02*	39.66*	59.22*	39.90*	68.24	45.41	69.35	46.60*
CS-rand	58.43	38.56	57.95	37.49	68.53	44.70	69.26	45.27

- SA, WA, CS contains more translation signals than BLOOM-560m
- WA provides comparable translation signals to LLMs compared with SA
- CS may provide some translation signals sometimes

Experiments — Training a small model

Tencent AI Lab

COMET and BLEURT may be unreliable to compare weak translation systems

	ZH-EN		EN-ZH	
	COMET	BLEURT	COMET	BLEURT
SA	38.07	18.47	32.54	5.16
SA-rand	23.96	8.48	24.19	2.82
WA	35.96	16.36	41.22	3.73
WA-rand	30.35	6.25	33.29	2.45
CS	39.40	18.71	37.10	6.39
CS-rand	37.45	17.63	37.91	6.49
random	36.15	3.54	31.18	0.73

Note: Random denotes randomly sample another source sentence as the translation

Experiments — Training a small model

Tencent AI Lab

Evaluation by conditional **perplexity** w.r.t P(e|f) defined by ICL

	ZH-EN			EN-ZH		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
SA	117.69	121.74	123.09	523.11*	525.32*	527.93*
SA-rand	110.85	109.13	109.00	–	–	–
WA	80.58*	81.21*	81.30*	216.33*	212.53*	212.10*
WA-rand	154.89	151.66	150.60	375.24	363.67	363.34
CS	129.82	131.35	132.37	270.67*	268.45	273.18
CS-rand	112.20	109.21	108.21	281.94	269.61	266.18

- WA achieves lower PPL than both SA and CS
- Exception: SA-random is better than SA.
  - Does it mean SA-random contain more translation signals than SA?

Experiments — Training a small model

Tencent AI Lab

Evaluation by conditional **perplexity** w.r.t P(e|f) defined by ICL

	ZH-EN				EN-ZH			
	target	1-shot	3-shot	5-shot	target	1-shot	3-shot	5-shot
SA	149.09	117.69	121.74	123.09	1303.59	523.11*	525.32*	527.93*
SA-rand	85.50	110.85	109.13	109.00	–	–	–	–
WA	115.72	80.58*	81.21*	81.30*	346.65	216.33*	212.53*	212.10*
WA-rand	130.63	154.89	151.66	150.60	489.26	375.24	363.67	363.34
CS	138.36	129.82	131.35	132.37	343.39	270.67*	268.45	273.18
CS-rand	91.34	112.20	109.21	108.21	351.53	281.94	269.61	266.18

Target column denotes the perplexity w.r.t the target language model P(e)
Underline "_" denotes the translation signal emerges

- SA-rand does not capture translation capability due to its limited size
  - P(e|f) for SA-random does not take into account f at all.
  - P(e|f) for SA-random is reduced to a target language model P(e).

Experiments — Purified data

Tencent AI Lab

Purified data: the other data **excluding** SA, WA and CS

# step	ZH-EN				EN-ZH			
	target	1-shot	3-shot	5-shot	target	1-shot	3-shot	5-shot
4.5k	145.54	155.12	148.94	147.00	562.75	804.49	727.77	711.32
7.5k	114.55	141.42	132.69	129.77	450.33	517.05	477.31	511.61
10k	66.13	90.49	83.75	81.87	242.68	200.21	182.51	179.54

Underline "_" denotes the translation signal emerges

- Surprisingly, it is possible to acquire translation signal by learning from purified data, although it is more difficult compared with learning from SA, WA or CS.
- Why learning from purified data enables translation capability?

Why purified data enables translation capability?

Tencent AI Lab

- Token Sharing in the data
  - Some common tokens such as numerical digits are shared across different languages in the training corpus

Sharing	Target	1-shot	3-shot	5-shot
✓	242.68	200.21	182.51	179.54
✗	410.69	734.54	512.7	432.36

Beijing will procure RMB 80 million in social organization services. Beijing News (新京报), January 28, 2013.

据航空数据提供商睿思誉的数据，中国航空公司从波音公司订购了至少209架737 机型，预计2024年中国航空公司将接收80架飞机。

W/ Sharing Tokens

Beijing will procure RMB eighty million in social organization services. Beijing News (新京报), January twenty-eight, two zero one three.

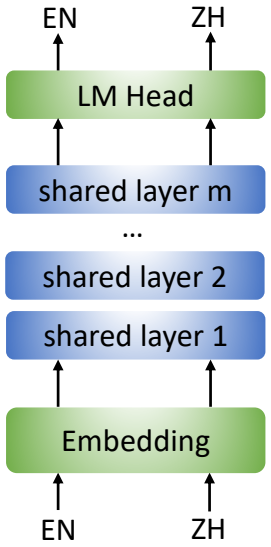
据航空数据提供商睿思誉的数据，中国航空公司从波音公司订购了至少二百零九架七三七 机型，预计二零零四年中国航空公司将接收八十架飞机。

W/o Sharing Tokens

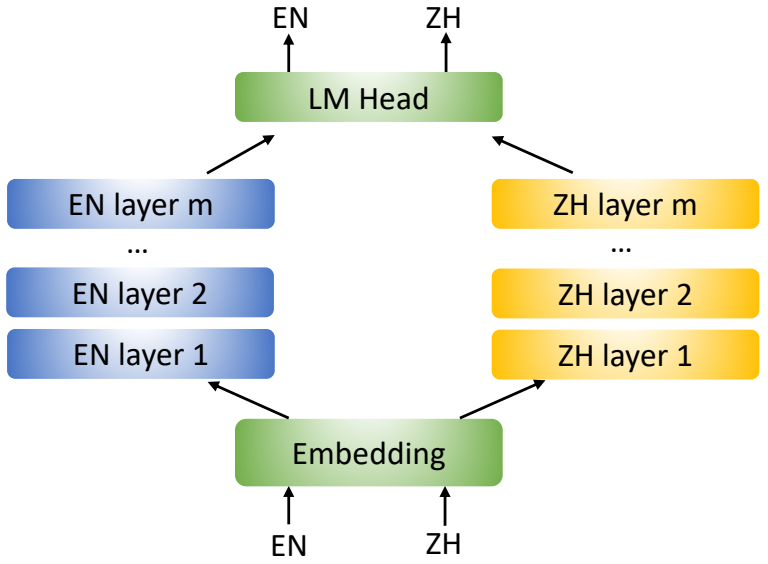
Why purified data enables translation capability?

- Parameter Sharing in LLMs
  - The dense parameters in the model are shared across different languages during the training process

Sharing	Target	1-shot	3-shot	5-shot
✓	242.68	200.21	182.51	179.54
✗	241.83	304.03	290.31	288.03



W/ Sharing Parameters



W/O Sharing Parameters

Summary

Tencent AI Lab

- Word-Alignment data provides comparable or sometimes superior translation signals to LLMs compared with sentence-alignment data
- Code-Switch data may also provide modest translation signals to LLMs
- Purified data may boost the translation capability of LLMs through common tokens (*e.g.* numeric digits) and sharing parameters in the architecture of LLMs across different languages

Future Work

Tencent AI Lab

- This work explores why LLMs enable translation capability but **how** LLMs learn to translate?
- It is **promising** to collect word-aligned data to boost the translation capability of LLMs
  - Sentence-level parallel data is limited especially for **low-resource** translation tasks

Tencent AI Lab

Thanks