NLP 南京大学 自然语言处理研究组 Natural Language Processing Group

南京大学 NANJING UNIVERSITY

Analyzing Multilingual Machine Translation Ability of Large Language Models

Shujian Huang and Wenhao Zhu
National Key Lab of Novel Software Technology
Department of Computer Science and Technology, Nanjing University

Outline

Part 1: Introduction

Part 2: Benchmarking LLM's Multilingual Translation Ability

Part 3: Investigating In-context Learning in Machine Translation

Part 4: Discussion

Part 5: Conclusion

Part 1: Introduction

Language Modeling

- **Language modeling aims at predicting the probability of the next token based on the prefix:**

$$p(w_t|w_1, \dots, w_{t-1})$$

- **Transformer has become the backbone of LLMs:**
  - Encoder-Decoder/Decoder Only
  - Scales Up

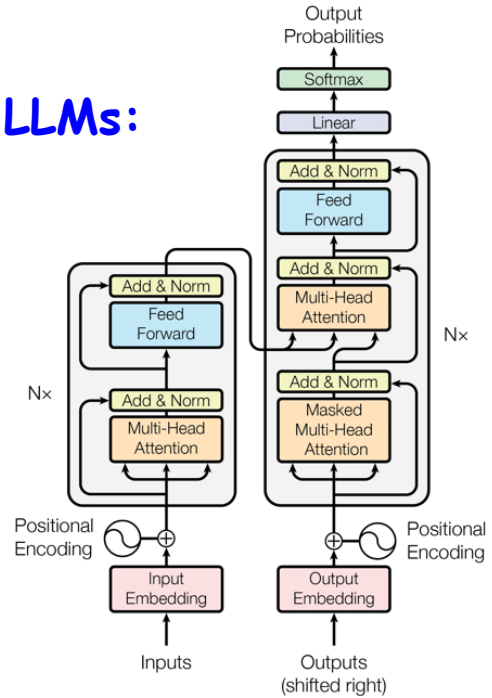


Image from: Attention is All you need. Vaswani et al., 2017.

Large Language Models

- More and more efforts have been put into building LLMs.

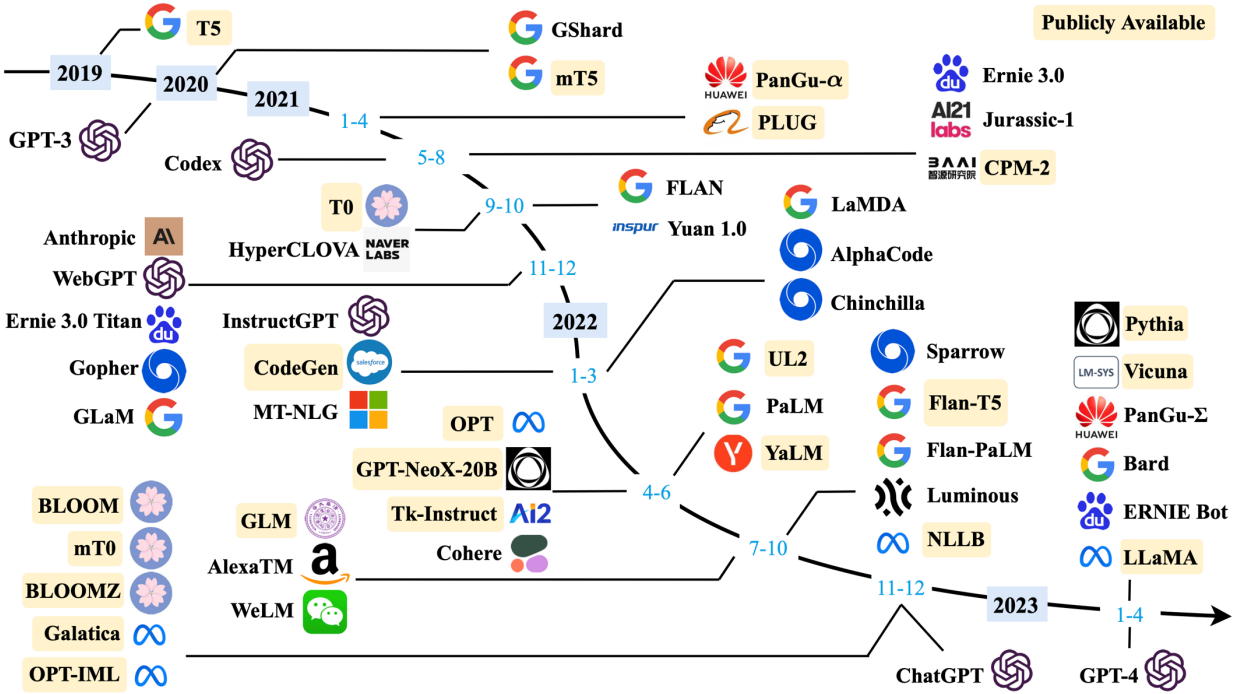


Image source: A Survey of Large Language Models. Zhao et al., arXiv'2023.

Scaling-up Language Model

- Increasing datasets and parameters leads to better language modeling for multiple languages, potentially useful for translation.

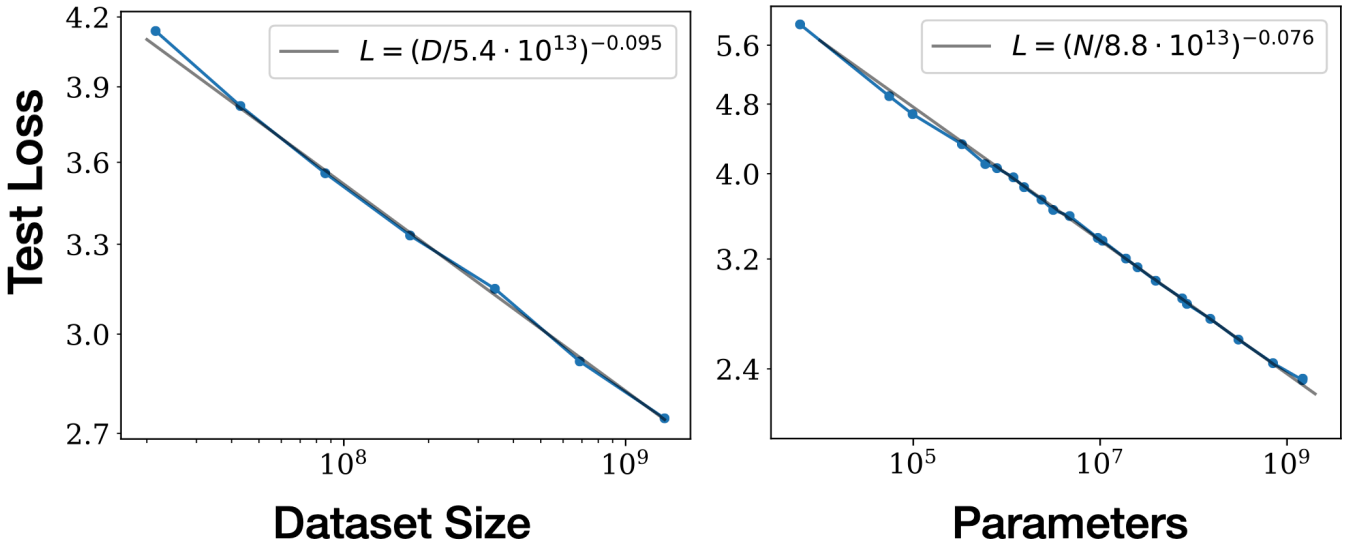
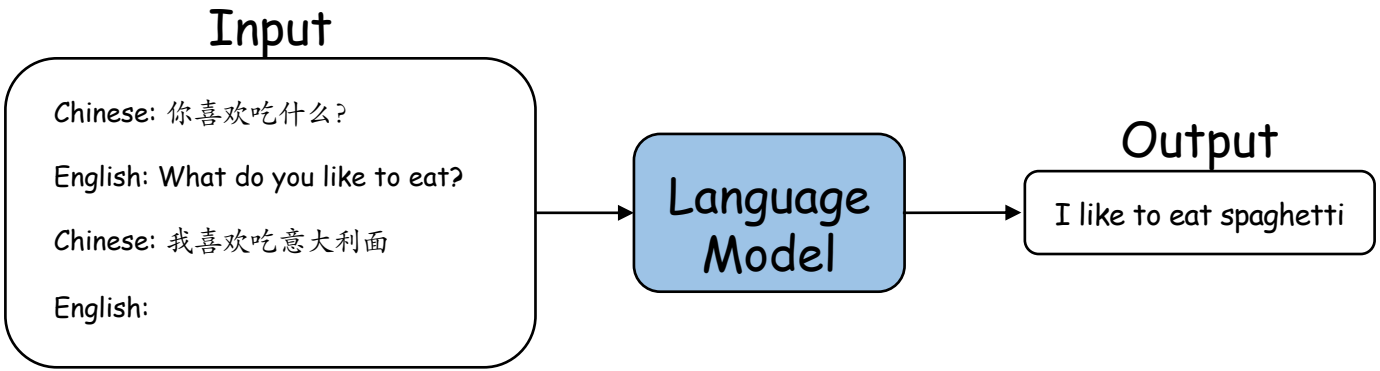


Image source: Scaling Laws for Neural Language Models. Kaplan et al., arXiv'2022.

Scaling-up Language Model

- **Scaling-up language model brings emergent ability, e.g., in-context learning (ICL), important for performing different tasks, including translation.**

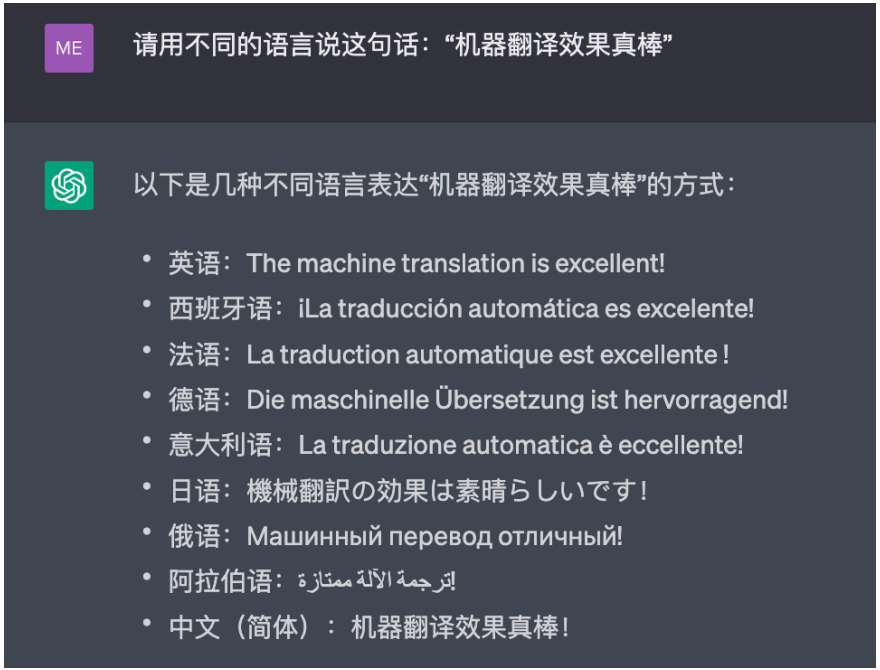


"In-context learning" allows language models to learn tasks with a few given exemplars without updating any parameter.

Emergent Abilities of Large Language Models. Wei et al., TMLR'2022.

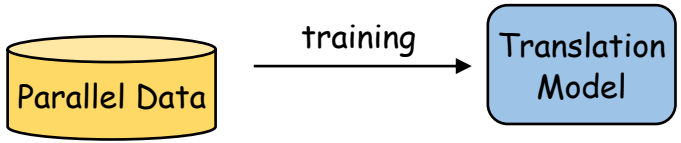
ChatGPT: A Representative LLM

- **ChatGPT is able to understand human instruction and has impressive multilingual ability.**

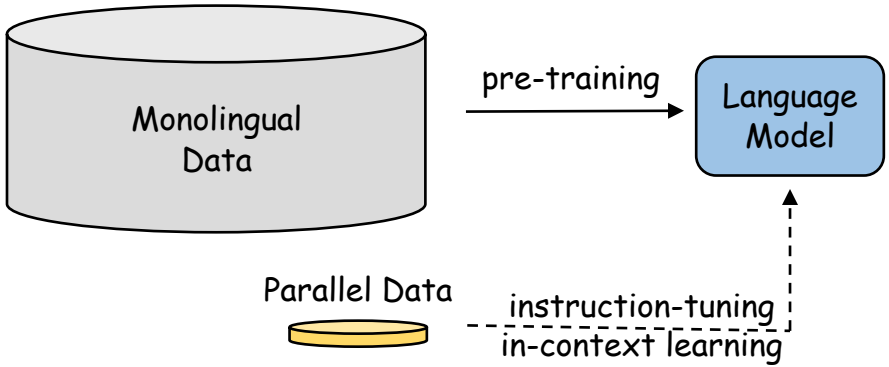


A Changing of Learning Paradigm

- **NMT: mainly from parallel data**



- **LLM: mainly from monolingual data**



What is the difference?

- **How well are LLMs in translating a massive number of languages?**
- **What affects LLMs' performance in translation?**
- **Our initial empirical study:**
  - Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. (arXiv 2304.04675)

Acknowledgement of Collaborators

- **Nanjing University**
  Wenhao Zhu, Shujian Huang, Jiajun Chen

- **Shanghai Jiaotong University**
  Hongyi Liu

- **Peking University**
  Qingxiu Dong

- **Shanghai AI Lab**
  Jingjing Xu, Lingpeng Kong (HKU)

- **University of California, Santa Barbara**
  Lei Li

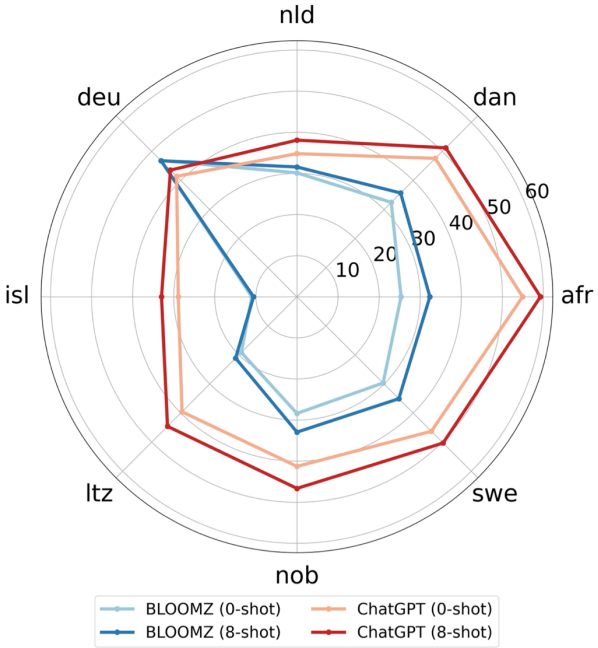
Part 2: Benchmarking LLM's Multilingual Translation Ability

Experiment Setup

- **LLMs**
  - Pre-trained: XGLM-7.5B (Lin et al., 2022), OPT-175B (Zhang et al. 2022)
  - Instruction-tuned: BLOOMZ-7.1B (Muennighoff et al., 2022), ChatGPT
- **ICL strategy**
  - ICL exemplars: eight randomly-picked translation pairs
  - ICL template: <X>=<Y>
- **Supervised baseline**
  - M2M-12B (Fan et al., 2020)、NLLB-1.3B
- **Test Dataset: Flores-101**
  - translating from English to 101 languages
  - covering a long tail of low-resource languages

Benchmarking Results

- **We evaluate all LLMs with ICL exemplars**
  - Even instruction-tuned LLMs (BLOOMZ, ChatGPT) can still benefit from in-context learning.



Benchmarking Results

- **OPT shows surprising multi-lingual performance even if it is not particularly optimized on multilingual data.**

Language Family	Direction	Pre-trained		Instruction-tuned		Supervised	
		XGLM-7.5B	OPT-175B	BLOOMZ-7.1B	ChatGPT	M2M-12B	NLLB-1.3B
Indo-European-Germanic (8)	X⇒En	18.54	34.65	30.77	45.83	42.72	**46.54**
	En⇒X	9.16	18.89	6.10	36.34	37.30	**38.47**
Indo-European-Romance (8)	X⇒En	31.11	38.93	**57.23**	45.68	42.33	46.33
	En⇒X	21.95	24.30	36.00	41.35	42.98	**43.48**
Indo-European-Slavic (12)	X⇒En	13.20	20.83	19.40	**39.27**	35.87	39.23
	En⇒X	6.40	8.18	2.50	32.61	35.01	**36.56**
Indo-European-Indo-Aryan (10)	X⇒En	8.68	1.20	**52.28**	25.32	19.59	40.75
	En⇒X	4.76	0.14	**47.97**	16.50	16.15	34.04
Indo-European-Other (11)	X⇒En	7.32	7.80	10.89	29.54	27.57	**39.87**
	En⇒X	4.51	3.10	2.10	22.81	24.31	**34.91**
Austronesian (6)	X⇒En	16.19	25.60	31.23	39.95	38.41	**45.41**
	En⇒X	10.01	10.68	16.82	30.17	32.64	**37.17**
Atlantic-Congo (14)	X⇒En	6.67	9.17	32.01	19.86	16.96	**32.20**
	En⇒X	2.52	1.60	7.20	8.91	10.71	**21.99**
Afro-Asiatic (6)	X⇒En	6.70	5.93	15.07	20.84	15.50	**39.43**
	En⇒X	2.07	1.40	4.39	13.57	12.24	**31.50**
Turkic (5)	X⇒En	7.43	7.89	6.26	24.64	13.07	**32.92**
	En⇒X	3.48	2.58	1.18	17.13	13.83	**30.17**
Dravidian (4)	X⇒En	8.04	0.89	**53.22**	20.26	14.01	39.07
	En⇒X	5.30	0.02	**57.50**	12.34	9.47	37.33
Sino-Tibetan (3)	X⇒En	9.35	9.32	**34.64**	21.36	17.13	30.88
	En⇒X	10.14	2.57	**38.78**	19.92	16.13	16.85
Other (14)	X⇒En	9.71	10.10	11.58	25.59	27.57	**30.88**
	En⇒X	8.42	3.82	5.37	20.26	25.16	**28.54**

bold: highest score of all models.
underline: highest score in LLMs.

Benchmarking Results

- OPT shows surprising multi-lingual performance even if it is not particularly optimized on multilingual data.
- Instruction-tuned LLMs often achieve better translation performance than pre-trained LLMs.

Language Family	Direction	Pre-trained		Instruction-tuned		Supervised	
		XGLM-7.5B	OPT-175B	BLOOMZ-7.1B	ChatGPT	M2M-12B	NLLB-1.3B
Indo-European-Germanic (8)	X⇒En	18.54	34.65	30.77	45.83	42.72	**46.54**
	En⇒X	9.16	18.89	6.10	36.34	37.30	**38.47**
Indo-European-Romance (8)	X⇒En	31.11	38.93	**57.23**	45.68	42.33	46.33
	En⇒X	21.95	24.30	36.00	41.35	42.98	**43.48**
Indo-European-Slavic (12)	X⇒En	13.20	20.83	19.40	**39.27**	35.87	39.23
	En⇒X	6.40	8.18	2.50	32.61	35.01	**36.56**
Indo-European-Indo-Aryan (10)	X⇒En	8.68	1.20	**52.28**	25.32	19.59	40.75
	En⇒X	4.76	0.14	**47.97**	16.50	16.15	34.04
Indo-European-Other (11)	X⇒En	7.32	7.80	10.89	29.54	27.57	**39.87**
	En⇒X	4.51	3.10	2.10	22.81	24.31	**34.91**
Austronesian (6)	X⇒En	16.19	25.60	31.23	39.95	38.41	**45.41**
	En⇒X	10.01	10.68	16.82	30.17	32.64	**37.17**
Atlantic-Congo (14)	X⇒En	6.67	9.17	32.01	19.86	16.96	**32.20**
	En⇒X	2.52	1.60	7.20	8.91	10.71	**21.99**
Afro-Asiatic (6)	X⇒En	6.70	5.93	15.07	20.84	15.50	**39.43**
	En⇒X	2.07	1.40	4.39	13.57	12.24	**31.50**
Turkic (5)	X⇒En	7.43	7.89	6.26	24.64	13.07	**32.92**
	En⇒X	3.48	2.58	1.18	17.13	13.83	**30.17**
Dravidian (4)	X⇒En	8.04	0.89	**53.22**	20.26	14.01	39.07
	En⇒X	5.30	0.02	**57.50**	12.34	9.47	37.33
Sino-Tibetan (3)	X⇒En	9.35	9.32	**34.64**	21.36	17.13	30.88
	En⇒X	10.14	2.57	**38.78**	19.92	16.13	16.85
Other (14)	X⇒En	9.71	10.10	11.58	25.59	27.57	**30.88**
	En⇒X	8.42	3.82	5.37	20.26	25.16	**28.54**

bold: highest score of all models.
underline: highest score in LLMs.

Benchmarking Results

- **BLOOMZ outperforms supervised baseline on seven groups of translation directions**

Language Family	Direction	Pre-trained		Instruction-tuned		Supervised	
		XGLM-7.5B	OPT-175B	BLOOMZ-7.1B	ChatGPT	M2M-12B	NLLB-1.3B
Indo-European-Germanic (8)	X⇒En	18.54	34.65	30.77	45.83	42.72	**46.54**
	En⇒X	9.16	18.89	6.10	36.34	37.30	**38.47**
Indo-European-Romance (8)	X⇒En	31.11	38.93	**57.23**	45.68	42.33	46.33
	En⇒X	21.95	24.30	36.00	41.35	42.98	**43.48**
Indo-European-Slavic (12)	X⇒En	13.20	20.83	19.40	**39.27**	35.87	39.23
	En⇒X	6.40	8.18	2.50	32.61	35.01	**36.56**
Indo-European-Indo-Aryan (10)	X⇒En	8.68	1.20	**52.28**	25.32	19.59	40.75
	En⇒X	4.76	0.14	**47.97**	16.50	16.15	34.04
Indo-European-Other (11)	X⇒En	7.32	7.80	10.89	29.54	27.57	**39.87**
	En⇒X	4.51	3.10	2.10	22.81	24.31	**34.91**
Austronesian (6)	X⇒En	16.19	25.60	31.23	39.95	38.41	**45.41**
	En⇒X	10.01	10.68	16.82	30.17	32.64	**37.17**
Atlantic-Congo (14)	X⇒En	6.67	9.17	32.01	19.86	16.96	**32.20**
	En⇒X	2.52	1.60	7.20	8.91	10.71	**21.99**
Afro-Asiatic (6)	X⇒En	6.70	5.93	15.07	20.84	15.50	**39.43**
	En⇒X	2.07	1.40	4.39	13.57	12.24	**31.50**
Turkic (5)	X⇒En	7.43	7.89	6.26	24.64	13.07	**32.92**
	En⇒X	3.48	2.58	1.18	17.13	13.83	**30.17**
Dravidian (4)	X⇒En	8.04	0.89	**53.22**	20.26	14.01	39.07
	En⇒X	5.30	0.02	**57.50**	12.34	9.47	37.33
Sino-Tibetan (3)	X⇒En	9.35	9.32	**34.64**	21.36	17.13	30.88
	En⇒X	10.14	2.57	**38.78**	19.92	16.13	16.85
Other (14)	X⇒En	9.71	10.10	11.58	25.59	27.57	**30.88**
	En⇒X	8.42	3.82	5.37	20.26	25.16	**28.54**

bold: highest score of all models.
underline: highest score in LLMs.

Benchmarking Results

- **BLOOMZ outperforms supervised baseline on seven groups of translation directions**
- **ChatGPT is the best translator among evaluated LLMs.**

Language Family	Direction	Pre-trained		Instruction-tuned		Supervised	
		XGLM-7.5B	OPT-175B	BLOOMZ-7.1B	ChatGPT	M2M-12B	NLLB-1.3B
Indo-European-Germanic (8)	X⇒En	18.54	34.65	30.77	45.83	42.72	**46.54**
	En⇒X	9.16	18.89	6.10	36.34	37.30	**38.47**
Indo-European-Romance (8)	X⇒En	31.11	38.93	**57.23**	45.68	42.33	46.33
	En⇒X	21.95	24.30	36.00	41.35	42.98	**43.48**
Indo-European-Slavic (12)	X⇒En	13.20	20.83	19.40	**39.27**	35.87	39.23
	En⇒X	6.40	8.18	2.50	32.61	35.01	**36.56**
Indo-European-Indo-Aryan (10)	X⇒En	8.68	1.20	**52.28**	25.32	19.59	40.75
	En⇒X	4.76	0.14	**47.97**	16.50	16.15	34.04
Indo-European-Other (11)	X⇒En	7.32	7.80	10.89	29.54	27.57	**39.87**
	En⇒X	4.51	3.10	2.10	22.81	24.31	**34.91**
Austronesian (6)	X⇒En	16.19	25.60	31.23	39.95	38.41	**45.41**
	En⇒X	10.01	10.68	16.82	30.17	32.64	**37.17**
Atlantic-Congo (14)	X⇒En	6.67	9.17	32.01	19.86	16.96	**32.20**
	En⇒X	2.52	1.60	7.20	8.91	10.71	**21.99**
Afro-Asiatic (6)	X⇒En	6.70	5.93	15.07	20.84	15.50	**39.43**
	En⇒X	2.07	1.40	4.39	13.57	12.24	**31.50**
Turkic (5)	X⇒En	7.43	7.89	6.26	24.64	13.07	**32.92**
	En⇒X	3.48	2.58	1.18	17.13	13.83	**30.17**
Dravidian (4)	X⇒En	8.04	0.89	**53.22**	20.26	14.01	39.07
	En⇒X	5.30	0.02	**57.50**	12.34	9.47	37.33
Sino-Tibetan (3)	X⇒En	9.35	9.32	**34.64**	21.36	17.13	30.88
	En⇒X	10.14	2.57	**38.78**	19.92	16.13	16.85
Other (14)	X⇒En	9.71	10.10	11.58	25.59	27.57	**30.88**
	En⇒X	8.42	3.82	5.37	20.26	25.16	**28.54**

bold: highest score of all models.
underline: highest score in LLMs.

Benchmarking Results

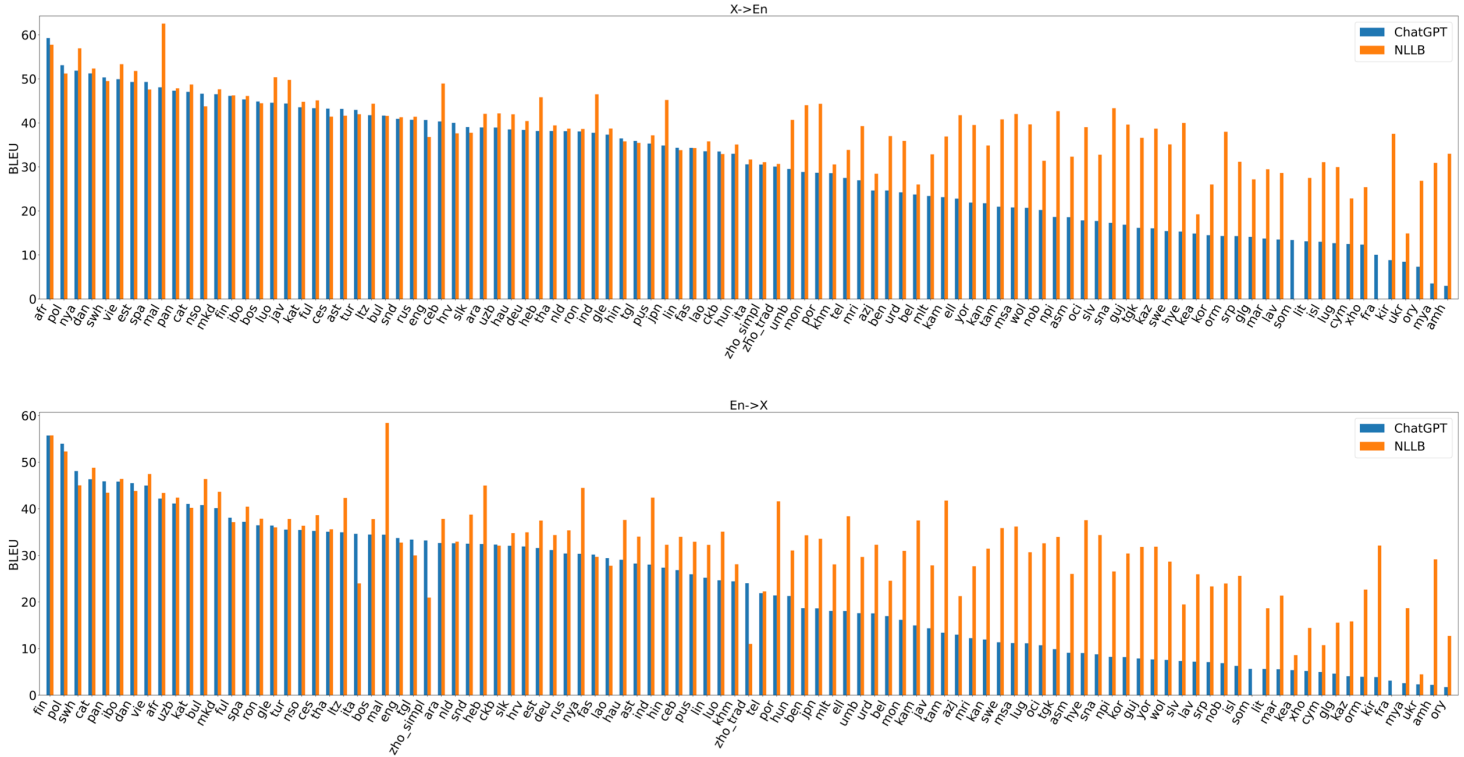
- BLOOMZ outperforms supervised baseline on seven groups of translation directions
- ChatGPT is the best translator among evaluated LLMs.
- LLMs perform better on translating into English than from translating from English.

Language Family	Direction	Pre-trained		Instruction-tuned		Supervised	
		XGLM-7.5B	OPT-175B	BLOOMZ-7.1B	ChatGPT	M2M-12B	NLLB-1.3B
Indo-European-Germanic (8)	X⇒En	18.54	34.65	30.77	45.83	42.72	**46.54**
	En⇒X	9.16	18.89	6.10	36.34	37.30	**38.47**
Indo-European-Romance (8)	X⇒En	31.11	38.93	**57.23**	45.68	42.33	46.33
	En⇒X	21.95	24.30	36.00	41.35	42.98	**43.48**
Indo-European-Slavic (12)	X⇒En	13.20	20.83	19.40	**39.27**	35.87	39.23
	En⇒X	6.40	8.18	2.50	32.61	35.01	**36.56**
Indo-European-Indo-Aryan (10)	X⇒En	8.68	1.20	**52.28**	25.32	19.59	40.75
	En⇒X	4.76	0.14	**47.97**	16.50	16.15	34.04
Indo-European-Other (11)	X⇒En	7.32	7.80	10.89	29.54	27.57	**39.87**
	En⇒X	4.51	3.10	2.10	22.81	24.31	**34.91**
Austronesian (6)	X⇒En	16.19	25.60	31.23	39.95	38.41	**45.41**
	En⇒X	10.01	10.68	16.82	30.17	32.64	**37.17**
Atlantic-Congo (14)	X⇒En	6.67	9.17	32.01	19.86	16.96	**32.20**
	En⇒X	2.52	1.60	7.20	8.91	10.71	**21.99**
Afro-Asiatic (6)	X⇒En	6.70	5.93	15.07	20.84	15.50	**39.43**
	En⇒X	2.07	1.40	4.39	13.57	12.24	**31.50**
Turkic (5)	X⇒En	7.43	7.89	6.26	24.64	13.07	**32.92**
	En⇒X	3.48	2.58	1.18	17.13	13.83	**30.17**
Dravidian (4)	X⇒En	8.04	0.89	**53.22**	20.26	14.01	39.07
	En⇒X	5.30	0.02	**57.50**	12.34	9.47	37.33
Sino-Tibetan (3)	X⇒En	9.35	9.32	**34.64**	21.36	17.13	30.88
	En⇒X	10.14	2.57	**38.78**	19.92	16.13	16.85
Other (14)	X⇒En	9.71	10.10	11.58	25.59	27.57	**30.88**
	En⇒X	8.42	3.82	5.37	20.26	25.16	**28.54**

bold: highest score of all models.
underline: highest score in LLMs.

Benchmarking Results

- LLMs still lag behind the strong supervised baseline, especially on low-resource languages.



Looking into Sino-Tibetan Languages

Language Family	Language	X-En					
		XGLM-7.5B	OPT-175B	BLOOMZ-7.1B	ChatGPT	M2M-12B	NLLB-1.3B
Sino-Tibetan (3)	mya	15.07	0.18	0.45	3.50	8.02	30.90
	zho_simpl	6.91	15.44	52.24	30.52	26.24	31.07
	zho_trad	6.06	12.36	51.24	30.05	-	30.67
		En-X					
		XGLM-7.5B	OPT-175B	BLOOMZ-7.1B	ChatGPT	M2M-12B	NLLB-1.3B
		9.60	0.02	0.04	2.57	7.28	18.66
		15.21	3.46	60.16	33.19	24.98	20.93
		5.63	4.22	56.13	24.01	-	10.97

- Translating Mya is hard for most systems (low-resource).
- Extremely high for BLOOMZ (later, overestimated due to data leakage)
- Extremely low for XGLM/OPT (later, not well activated)
- ChatGPT is still good, comparing to supervised baselines

LLMs' Typical Translation Errors

Type I: Off-target translation	
Source	Prime Minister Stephen Harper has agreed to send the government's 'Clean Air Act' to an all-party committee for review, before its second reading, after Tuesday's 25 minute meeting with NDP leader Jack Layton at the PMO.
Hypothesis	Baş nazir Stephen Harper, NDP lideri Jack Layton ile PMO'daki 25 dakikalık görüşmesinin ardından hükümetin 'Temiz Hava Yasası'nı ikinci okumadan önce tüm parti komitesine gözden geçirmesi için göndermeyi kabul etti.
Reference	Baş Nazir Stiven Harper Layihe İdareetme Ofisinde (LIO) Yeni Demokratik Partiya (YDP) lideri Cek Layton ile çerşenbe axşamı 25 deqiqelik iclasdan sonra ikinci oxunuşdan evvel hökumetin Ţemiz Hava Aktınıñezerden keçirilmesi üçün bütün partiyaların üzv olduğu komiteye göndermeyi qebul etdi
Error	ChatGPT fails to translate English into turkish, but into Azerbaijani.

LLMs' Typical Translation Errors

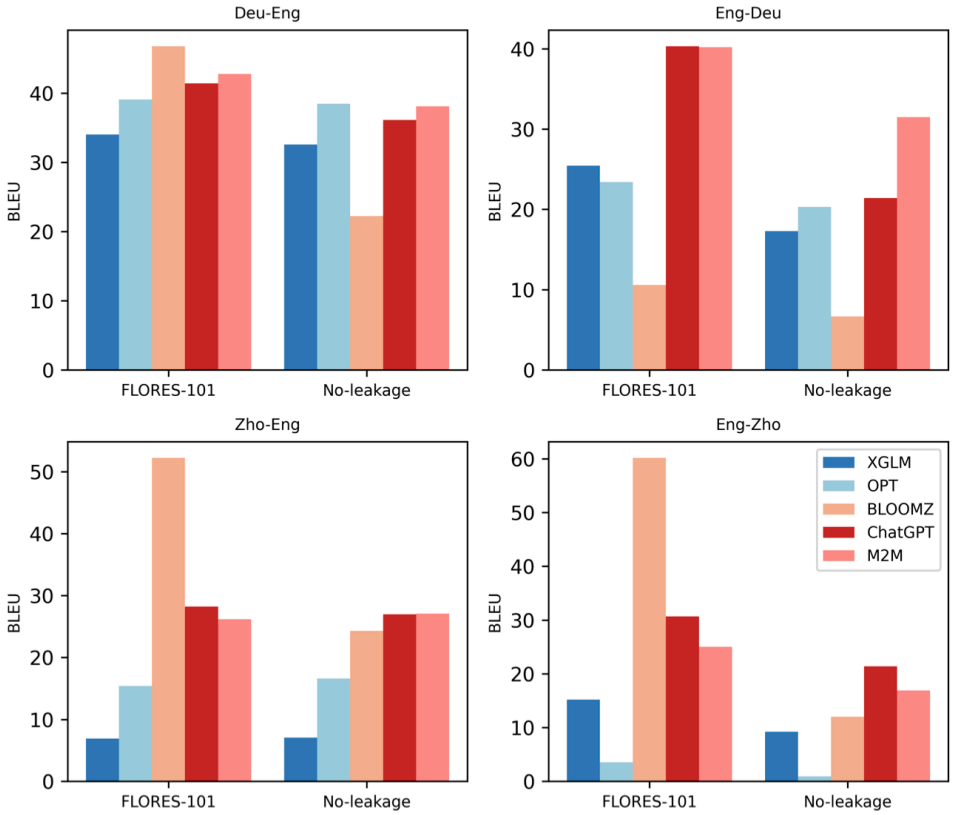
Type II: Hallucination	
Source	Huwobe illimu kimiya be fasaha jodi hefti do kujjeji ko jomirawo tagi ha asamaji do mari alaka be luttukal ko taginga ha asamaji fu bana no mari alaka be kala ko be tagi andigel pat.
Hypothesis	The study of chemistry and language is important in our modern world, and we must work hard to improve our knowledge in both areas. Let us all strive to learn and improve ourselves in these fields so that we can make a positive impact on society.
Reference	The scientists were able to conclude that the dark matter affect other dark matter in the same way regular matter does.
Error	ChatGPT generates highly pathological translations that are completely unrelated with the source sentence.

LLMs' Typical Translation Errors

Type III: Monotonic translation	
Source	They found the Sun operated on the same basic principles as other stars: The activity of all stars in the system was found to be driven by their luminosity, their rotation, and nothing else.
Hypothesis	他们发现太阳按照同样的基本原理运作，就像其他星体一样：系统中所有星体的活动均由它们的亮度、旋转驱动，且仅有这些因素
Reference	他们发现，太阳和其他恒星的运行原理是一样的：星系中所有恒星的活跃度都完全决定于它们的光度和自转
Error	ChatGPT translates English sentence word-by-word, lacking effective word-reordering.

Data Leakage Issue

- **Separate evaluation with recent news (no-leakage):**
  - XGLM and OPT perform consistently.
  - BLOOMZ drops dramatically (20-40).
    - *instruction-tuning on Flores-200
  - ChatGPT drops on En-De (10+).
  - M2M perform consistently.
- **Evaluating LLMs on public datasets have the risk of data leakage!**



Part 3: Investigating In-context Learning in Machine Translation

The analyses are conducted on XGLM-7.5B.

Findings on In-context Template

- The translation performance varies greatly with different template.
- The best template for each direction is also different.

In-context Template	Deu-Eng	Eng-Deu	Rus-Eng	Eng-Rus	Rus-Deu	Deu-Rus	Average
reasonable templates:							
<X>=<Y>	37.37	**26.49**	29.66	22.25	17.66	**17.31**	**25.12**
<X> \n Translate from [SRC] to [TGT]: \n <Y>	37.95	26.29	29.83	20.61	17.56	15.93	24.70
<X> \n Translate to [TGT]: \n <Y>	37.69	25.84	**29.96**	19.61	17.44	16.48	24.50
<X> \n [TGT]: <Y>	29.94	17.99	25.22	16.29	12.28	11.71	18.91
<X> is equivalent to <Y>	23.00	4.21	17.76	9.44	8.14	9.84	12.07
<X>\n can be translated to\n <Y>	37.55	**26.49**	29.82	22.14	17.48	16.40	24.98
[SRC]: <X> \n [TGT]: <Y>	16.95	8.90	14.48	6.88	7.86	4.01	9.85

Findings on In-context Template

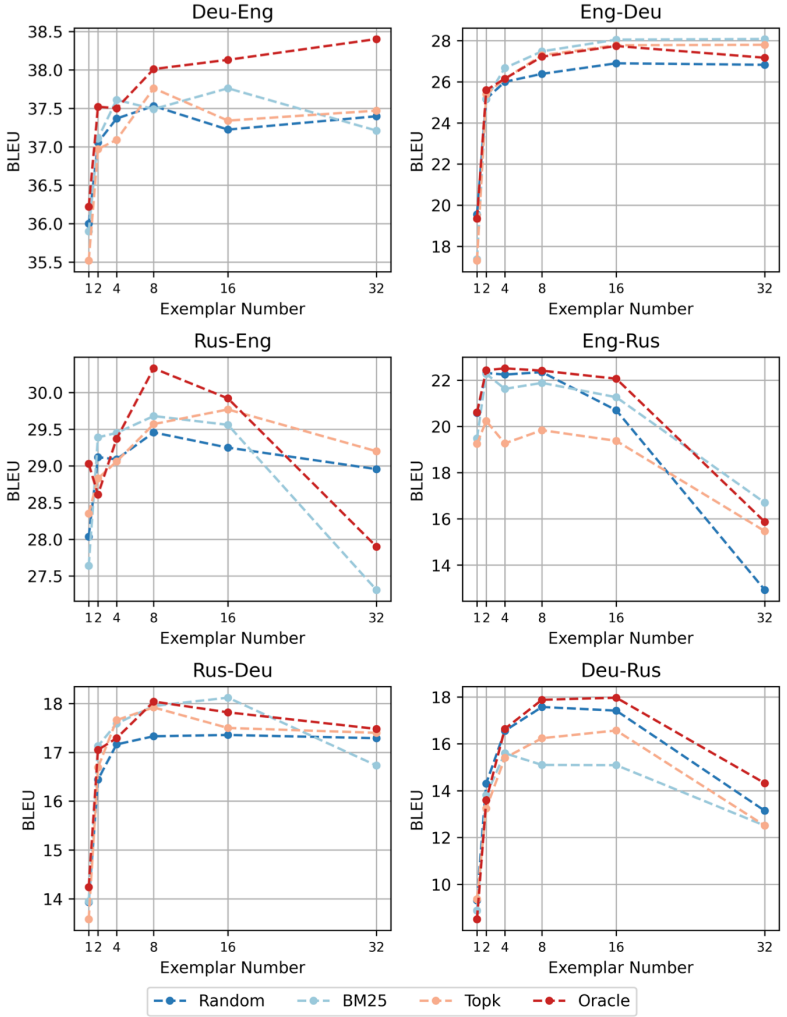
- Even unreasonable template can instruct LLM to generate decent translation.

In-context Template	Deu-Eng	Eng-Deu	Rus-Eng	Eng-Rus	Rus-Deu	Deu-Rus	Average
reasonable templates:							
<X>=<Y>	37.37	**26.49**	29.66	22.25	17.66	**17.31**	**25.12**
<X> \n Translate from [SRC] to [TGT]: \n <Y>	37.95	26.29	29.83	20.61	17.56	15.93	24.70
<X> \n Translate to [TGT]: \n <Y>	37.69	25.84	**29.96**	19.61	17.44	16.48	24.50
<X> \n [TGT]: <Y>	29.94	17.99	25.22	16.29	12.28	11.71	18.91
<X> is equivalent to <Y>	23.00	4.21	17.76	9.44	8.14	9.84	12.07
<X>\n can be translated to\n <Y>	37.55	**26.49**	29.82	22.14	17.48	16.40	24.98
[SRC]: <X> \n [TGT]: <Y>	16.95	8.90	14.48	6.88	7.86	4.01	9.85
unreasonable templates:							
<X>$<Y>	37.77	26.43	29.53	20.99	17.72	17.27	24.95
<X> \n Translate from [TGT] to [SRC]: \n <Y>	**38.18**	26.21	29.85	20.35	**17.75**	16.63	24.83
<X> \n Compile to [TGT]: \n <Y>	37.39	26.35	29.68	19.91	17.52	16.15	24.50
<X> \n [SRC]: <Y>	27.86	16.69	24.41	18.16	11.98	12.60	18.62
<X> is not equivalent to <Y>	23.50	3.92	16.90	7.80	8.06	9.23	11.57
<X> \n can be summarized as \n <Y>	37.46	26.24	29.42	**22.62**	17.68	17.15	25.10
[SRC]: <X> \n [SRC]: <Y>	19.03	8.21	15.96	6.37	7.57	4.40	10.26

Findings on In-context Exemplar

- **Four ways of selecting ICL exemplars**
  - Random
  - BM25 (sparse retrieval)
  - TopK (dense retrieval)
  - Oracle (selecting with reference)

- **Semantically-selected exemplars does not bring more benefits than randomly-picked exemplars.**



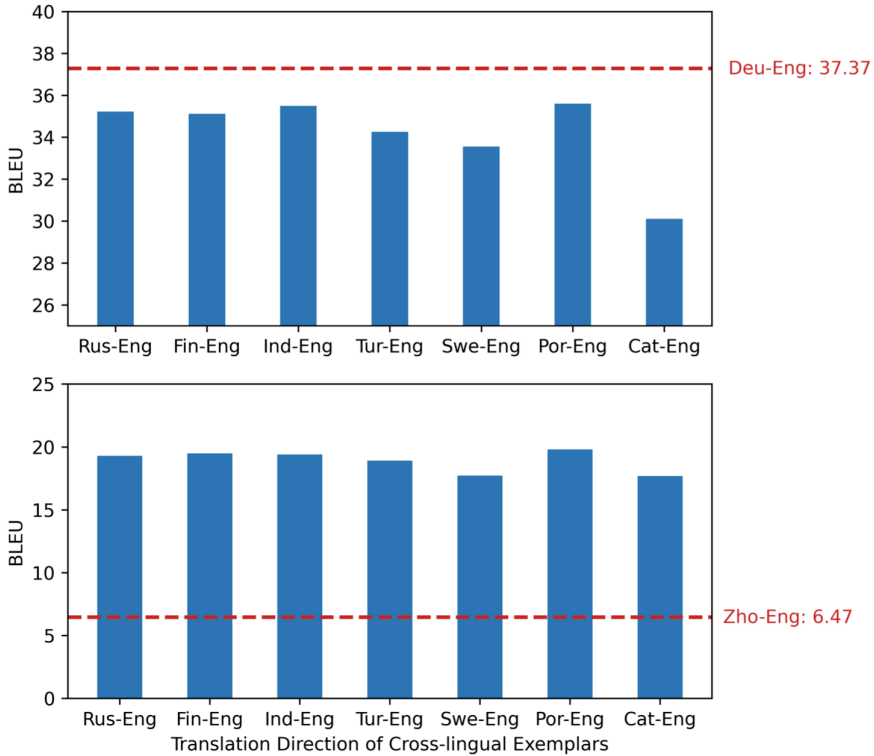
Findings on In-context Exemplar

- **ICL exemplars teach LLM the core feature of translation task.**
  - Keeping source and target sentence semantically consistent.
  - Adjusting translation granularity.
  - Translating case by case.

In-context Exemplars	Consistency	Granularity	Diversity	Deu-Eng	Eng-Deu
Mismatched Translation	✗	✓	✓	0.00	0.00
Word-level Translation	✓	✗	✓	25.10	5.84
Doc-level Translation	✓	✗	✓	8.01	2.05
Duplicated Translation	✓	✓	✗	35.12	19.66
Sent-level Translation	✓	✓	✓	37.37	26.49

Findings on In-context Exemplar

- **Using cross-lingual exemplars does not always cause worse performance.**



Activating the translation ability is also important.

Findings on In-context Exemplar

- **The exemplar in the tail of the prompt has more impact**
  - Reversing exemplars in the tail of the prompt consistently produced worse results compared to reversing in the head.

Rev ratio	Deu-Eng Head	Deu-Eng Tail	Eng-Deu Head	Eng-Deu Tail
0 / 8	37.37	37.37	26.49	26.49
1 / 8	37.74	36.05	26.75	23.96
2 / 8	37.29	36.79	26.89	24.66
3 / 8	36.82	35.67	26.44	24.34
4 / 8	36.60	35.18	26.23	22.17
5 / 8	35.61	31.93	25.58	17.47
6 / 8	30.49	20.71	22.42	8.73
7 / 8	14.60	5.36	12.51	3.19
8 / 8	3.42	3.42	3.10	3.10

Part 4: Discussion

Learn translation from unsupervised data

- Training an 8B LM, with no parallel data (Garcia et al.)
  - BLEURT (top)
  - BLEU (bottom)

Language	Examples
English	69,813,325
German	69,813,325
Chinese	33,172,846
Icelandic	250,416

Model	*en ↔ zh* *newstest21*		*en ↔ de* *newstest21*	
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	75.9	74.8
Bilingual LMs (Beam)	62.6	67.0	74.9	74.1
Bilingual LMs (MBR)	68.4	67.8	75.5	76.5
Trilingual LM (Beam)	65.3	65.3	74.5	74.4
Trilingual LM (MBR)	68.9	**68.3**	75.5	76.8

Model	*en ↔ is* *newstest21*	
Supervised models		
WMT'21 1st Place	**77.2**	**76.1**
WMT'21 2nd Place	74.3	72.3
WMT'21 3rd Place	74.3	70.4
Google Translate	76.8	71.1
Few-shot translation models		
PaLM	61.7	59.5
Bilingual LMs (MBR)	76.2	72.0

Models	*zh ↔ en* *newstest21*		*de ↔ en* *newstest21*		*is ↔ en* *newstest21*	
Supervised models						
WMT'21 1st Place	33.4	36.9	41.9	42	41.7	33.3
WMT'21 2nd Place	31.9	35.9	39.7	43.2	40	30.6
WMT'21 3rd Place	32.6	35.8	40	41.3	39.2	28.6
Google Translate	32.2	36.2	40.9	39.8	41.5	28.7
Few-shot translation models						
PaLM	25.8	29.6	38.8	32.9	19.1	16.87
Bilingual LM (MBR)	21.6	25.6	37.1	29.4	33.52	17.7
Bilingual LM (Beam)	20.4	29.2	35.5	32.8	36.2	19.2
Trilingual LM (MBR)	22.2	26.8	36.1	28.5	–	–
Trilingual LM (Beam)	20.45	25.5	36.2	31.8	–	–

Table source: The unreasonable effectiveness of few-shot learning for machine translation. Garcia et al. arXiv'2023.

Unbalanced capability across languages

- It is usually better at handling languages with richer resources.
  - Ratios come from XGLM paper (Lin et al., 2022)

Language	Ratio	En-X	X-En
Russian	6.0%	27.83	23.18
German	3.6%	34.04	25.44
Spanish	3.5%	27.98	23.82
Urdu	0.5%	19.31	13.63
Burmese	0.3%	15.07	9.60
Telugu	0.2%	17.22	12.25

- Up-sampled languages are indeed above average performance.

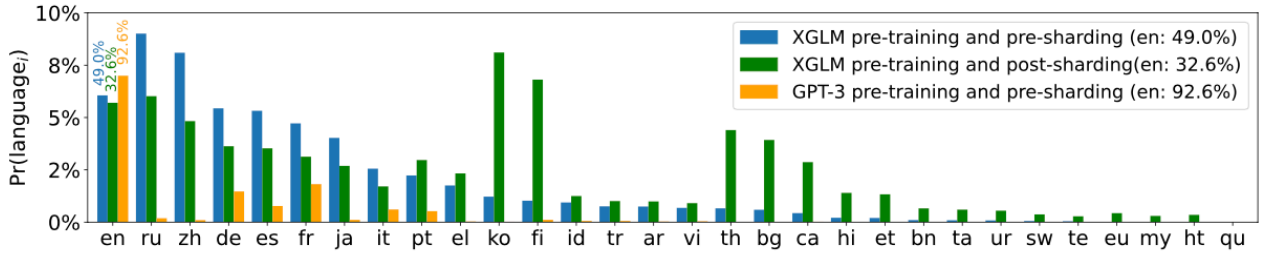


Figure source: Lin et al. Few-shot learning with multilingual generative language models. EMNLP'2022.

Part 5: Conclusion

Conclusion

- We evaluate the multilingual translation ability of several LLMs, including ChatGPT on 102 languages and 202 English-centric directions.
- Even the best-performed LLM (ChatGPT) still lags behind the strong multilingual supervised baseline (NLLB) in 83.33% translation directions.
- We find that LLMs exhibit some new working patterns when used for machine translation.

Acknowledgement of Collaborators （again）

- **Nanjing University**
  Wenhao Zhu, Shujian Huang, Jiajun Chen

- **Shanghai Jiaotong University**
  Hongyi Liu

- **Peking University**
  Qingxiu Dong

- **Shanghai AI Lab**
  Jingjing Xu, Lingpeng Kong (HKU)

- **University of California, Santa Barbara**
  Lei Li

Reference

- Few-shot learning with multilingual generative language models. Lin et al., EMNLP'2022.
- Cross-lingual Generalization through Multitask Finetuning. Muennighoff et al., arXiv'2022.
- OPT: Open Pre-trained Transformer Language Models. Zhang et al., arXiv'2022.
- Beyond English-Centric Multilingual Machine Translation. Fan et al., arXiv'2020
- No Language Left Behind: Scaling Human-Centered Machine Translation. NLLB Team, arXiv'2022

Thanks for Watching !