Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated remarkable potential in handling multilingual machine translation (MMT). In this paper, we systematically investigate the advantages and challenges of LLMs for MMT by answering two questions: 1) How well do LLMs perform in translating massive languages? 2) Which factors affect LLMs' performance in translation? We thoroughly evaluate eight popular LLMs, including ChatGPT and GPT-4. Our empirical results show that translation capabilities of LLMs are continually involving. GPT-4 has beat the strong supervised baseline NLLB in 40.91% of translation directions but still faces a large gap towards the commercial translation system like Google Translate, especially on low-resource languages. Through further analysis, we discover that LLMs exhibit new working patterns when used for MMT. First, LLM can acquire translation ability in a resource-efficient way and generate moderate translation even on zero-resource languages. Second, instruction semantics can surprisingly be ignored when given in-context exemplars. Third, cross-lingual exemplars can provide better task guidance for low-resource translation than exemplars in the same language pairs.

1 Introduction

With the increasing scale of parameters and training corpus, large language models (LLMs) have gained a universal ability to handle a variety of tasks via in-context learning (ICL, Brown et al. 2020), which allows language models to perform tasks with a few given exemplars and human-written instructions as context. One particular area where LLMs have shown outstanding potential is machine translation (MT). Previous studies have shown the surprising performance of LLMs on high-resource bilingual translation, such as English-German translation (Vilar et al., 2022; Zhang et al., 2022), even if these models are not particularly optimized on multilingual data.

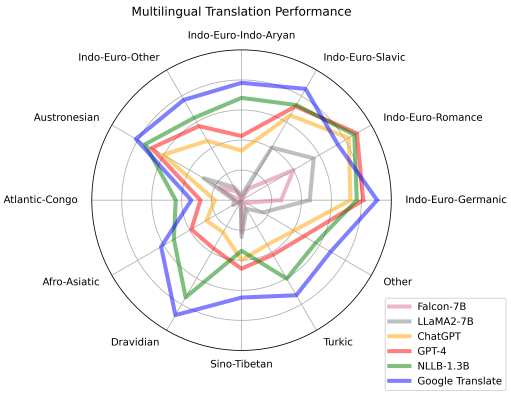

Figure 1: Multilingual translation performance (translating from English to non-English) of some popular LLMs and traditional supervised systems. LLMs have demonstrated great potential in multilingual machine translation.

However, the multilingual translation ability of LLMs remains under-explored. MMT is a challenging task that involves translating text among different languages and requires semantic alignment between languages (Fan et al., 2021; Costa-jussà et al., 2022; Yuan et al., 2023). It is also unclear that how LLM acquires translation ability and which factors affect LLM's translation ability.

In this paper, we follow ICL paradigm and focus on studying LLMs in multilingual machine translation by answering two questions: 1) *How LLMs perform MMT over massive languages?* 2) *Which factors affect the performance of LLMs?*

For the first question, we evaluate several popular LLMs: English-centric LLMs, including OPT (Zhang et al., 2022), LLaMA2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023) and multilingual LLMs, including XGLM (Lin et al., 2022), BLOOMZ (Scao et al., 2022), ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023). We consider 102 languages and 606 translation directions (202 English-centric directions, 202 French-centric directions and 202 Chinese-centric direc-

tions). Results show that the multilingual translation capabilities of LLMs are continually involving and GPT-4 reaches new performance height. Compared with the widely-used supervised MMT system NLLB (Costa-jussà et al., 2022), GPT-4 achieves higher performance on 40.91% English-centric translation directions. But compared with the commercial translation system (Google Translate), LLMs still have a long way to go, particularly when it comes to low-resource languages. French-centric and Chinese-centric translation are also more challenging for GPT-4 than English-centric translation, which further indicates its unbalanced capability across languages.

For the second question, we find some new working patterns. First, we discover that LLM can acquire translation ability in a resource-efficient way and generate moderate translation even on zero-resource languages. Second, LLMs are able to perform translation even with unreasonable instructions if in-context learning exemplars are given. However, if given mismatched translation pairs as in-context exemplars, LLMs fail to translate, which is similar to observations from concurrent studies (Wei et al., 2023). This shows the importance of exemplars in ICL for machine translation. Third, we find that cross-lingual translation pairs can be surprisingly good exemplars for low-resource translation, even better than exemplars in the same language.

The main contribution of this paper can be summarized below:

- We benchmark popular LLMs on MMT in 102 languages and 606 translation directions, covering English-centric, French-centric and Chinese-centric translation.
- We systematically compare the results of LLMs and three strong supervised baselines (M2M-100, NLLB, Google Translator) and reveal the gap between two translation paradigms.
- We find some new ICL working patterns of LLMs for MMT and discuss corresponding advantages and challenges.

2 Background

2.1 Large Language Models

Language modeling is a long-standing task in natural language processing (Bengio et al., 2000; Mikolov et al., 2010; Khandelwal et al., 2020), which is a task to predict the probability of the next token. Transformer (Vaswani et al., 2017) basically is the backbone of existing LLMs.

LLMs show great potential as a universal multi-task learner. Recently, Radford et al. (2019) find that a casual decoder-only language model can be a multi-task learner with merely unsupervised training corpus. Later, Kaplan et al. (2020) reveal the *scaling law* of LLM, indicating that when the scale of neural parameters and training data keeps increasing, LLM can be further strengthened. Wei et al. (2022b) show that scaling the language model also brings astonishing *emergent abilities*, e.g., in-context learning, which is only present in large models. Consequently, more and more efforts have been put into scaling-up language models (Brown et al., 2020; Hoffmann et al., 2022; Scao et al., 2022; Vilar et al., 2022; Ren et al., 2023). Among them, GPT-4 (OpenAI, 2023) and ChatGPT (OpenAI, 2022) are the most representative systems, which shows impressive results in various NLP tasks.

2.2 Emergent Ability: In-context Learning

In-context learning is one of the well-known emergent abilities (Brown et al., 2020; Dong et al., 2022), which enables LLM to learn target tasks according to the prompt without updating any parameters.

Specifically, the prompt is made up of in-context exemplars $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^{k}$ and in-context template $\mathcal{T}$. Exemplars are often picked from supervised data, where $\mathcal{Y}_i$ is the ground truth corresponding to the input sentence $\mathcal{X}_i$. Template $\mathcal{T}$ is usually a human-written instruction related to the target task. Wrapping exemplars with the template and concatenating them together produce the final prompt:

$$\mathcal{P} = \mathcal{T}(\mathcal{X}_1, \mathcal{Y}_1) \oplus \mathcal{T}(\mathcal{X}_2, \mathcal{Y}_2) \oplus \cdots \oplus \mathcal{T}(\mathcal{X}_k, \mathcal{Y}_k)$$

where $\oplus$ denotes the concatenation symbol, e.g., whitespace, line-break. During inference, LLM is able to generate the corresponding output $\mathcal{Y}$ of the test sample $\mathcal{X}$ under the guidance of the prompt:

$$\arg\max_{\mathcal{Y}} \; p(\mathcal{P} \oplus \mathcal{T}(\mathcal{X}, \mathcal{Y})) \qquad (1)$$

For label prediction tasks, the prediction $\mathcal{Y}$ can be obtained in one-step generation. For sequence generation tasks, e.g., machine translation, the prediction $\mathcal{Y}$ can be obtained through sampling strategies like greedy search and beam search.

Language Family	Direction	Translation Performance (BLEU / COMET)									
		XGLM-7.5B	OPT-175B	Falcon-7B	LLaMA2-7B	LLaMA2-7B-Chat	ChatGPT	GPT-4	M2M-12B	NLLB-1.3B	Google
Indo-Euro-Germanic (8)	X⇒Eng	18.54 / 70.09	34.65 / 83.71	27.37 / 67.40	37.28 / 84.73	34.82 / 84.25	45.83 / 89.05	48.51 / **89.48**	42.72 / 87.74	46.54 / 88.18	**51.16** / 89.36
	Eng⇒X	9.16 / 50.21	18.89 / 71.97	13.19 / 52.93	22.78 / 76.05	19.44 / 73.63	36.34 / 87.83	40.64 / 88.50	37.30 / 86.47	38.47 / 87.31	**45.27** / **89.05**
Indo-Euro-Romance (8)	X⇒Eng	31.11 / 79.67	38.93 / 87.75	34.06 / 84.40	41.10 / 88.10	37.84 / 87.80	45.68 / 89.61	**47.29** / **89.74**	42.33 / 88.31	46.33 / 88.99	35.69 / 89.66
	Eng⇒X	21.95 / 69.08	24.30 / 79.07	20.02 / 70.36	27.81 / 82.05	25.50 / 79.67	41.35 / **89.00**	**44.47** / 88.94	42.98 / 87.56	43.48 / 88.12	37.10 / 88.77
Indo-Euro-Slavic (12)	X⇒Eng	13.20 / 64.24	20.83 / 74.80	13.15 / 57.34	34.00 / 84.90	30.94 / 83.90	39.27 / 87.74	41.19 / 88.15	35.87 / 85.97	39.23 / 87.08	**43.61** / **88.18**
	Eng⇒X	6.40 / 43.28	8.18 / 54.45	4.34 / 35.73	20.24 / 76.30	16.14 / 69.75	32.61 / 87.90	36.06 / 89.15	35.01 / 86.43	36.56 / 88.74	**42.75** / **90.05**
Indo-Euro-Indo-Aryan (10)	X⇒Eng	8.68 / 63.93	1.20 / 49.37	1.40 / 45.22	6.68 / 62.63	4.29 / 60.29	25.32 / 84.14	37.30 / 87.79	17.53 / 69.66	40.75 / 88.80	**45.66** / **89.43**
	Eng⇒X	4.76 / 40.99	0.14 / 31.85	0.13 / 25.84	1.61 / 35.92	1.24 / 34.74	16.50 / 68.43	21.35 / 73.75	14.44 / 65.32	34.04 / 82.55	**39.04** / **82.78**
Indo-Euro-Other (11)	X⇒Eng	7.32 / 55.29	7.80 / 59.60	7.04 / 51.59	14.27 / 69.87	11.46 / 67.64	29.54 / 84.52	37.29 / 86.76	22.38 / 77.47	36.16 / 86.81	**41.68** / **88.29**
	Eng⇒X	4.51 / 40.60	3.10 / 40.04	3.38 / 34.64	5.00 / 44.09	4.83 / 43.73	22.81 / 77.33	28.45 / 80.94	19.71 / 74.90	31.65 / 85.82	**38.54** / **87.44**
Austronesian (6)	X⇒Eng	16.19 / 78.80	25.60 / 78.03	18.62 / 75.36	26.70 / 80.21	24.39 / 80.39	39.95 / 87.29	46.81 / 88.65	31.84 / 84.76	45.41 / 87.85	**50.68** / **88.89**
	Eng⇒X	10.01 / 73.14	10.68 / 64.97	8.56 / 60.89	14.59 / 74.80	13.29 / 74.88	30.17 / 86.36	34.66 / 87.68	27.03 / 86.83	37.17 / 88.82	**40.74** / **89.34**
Atlantic-Congo (14)	X⇒Eng	6.67 / 62.00	9.17 / 57.59	6.98 / 0.56	8.76 / 57.72	9.01 / 57.86	19.86 / 79.63	28.27 / 83.42	10.55 / 76.43	**32.20** / 84.00	23.55 / **85.44**
	Eng⇒X	2.52 / 54.93	1.60 / 34.15	1.89 / 0.34	2.45 / 34.17	3.09 / 38.13	8.91 / 75.26	13.70 / 77.79	6.53 / 75.79	**21.99** / 79.95	16.77 / **80.89**
Afro-Asiatic (6)	X⇒Eng	6.70 / 54.51	5.93 / 52.90	4.87 / 38.62	10.41 / 57.72	8.65 / 58.27	20.84 / 70.39	30.48 / 78.76	10.00 / 66.98	32.69 / 82.99	**36.14** / **84.47**
	Eng⇒X	2.07 / 41.48	1.40 / 41.86	1.40 / 27.64	3.22 / 43.04	3.07 / 43.39	13.57 / 67.60	19.36 / 75.56	7.83 / 68.86	26.08 / 82.84	**31.00** / **83.78**
Turkic (5)	X⇒Eng	7.43 / 61.69	7.89 / 62.47	4.15 / 33.11	9.51 / 65.95	8.88 / 66.15	24.64 / 84.04	31.73 / 86.90	10.25 / 58.52	32.92 / 87.51	**37.78** / **88.53**
	Eng⇒X	3.48 / 40.32	2.58 / 44.80	1.75 / 20.00	3.28 / 39.65	3.09 / 41.97	17.13 / 74.77	20.96 / 78.50	10.87 / 68.21	30.17 / 88.47	**36.54** / **89.38**
Dravidian (4)	X⇒Eng	8.04 / 61.95	0.89 / 44.01	1.18 / 24.29	2.65 / 53.17	1.52 / 52.95	20.26 / 82.00	33.10 / 86.91	10.26 / 63.77	39.07 / 88.42	**43.17** / **89.10**
	Eng⇒X	5.30 / 48.15	0.02 / 32.51	0.03 / 15.31	0.56 / 34.03	0.58 / 35.65	12.34 / 64.74	18.60 / 75.15	6.85 / 62.25	37.33 / 86.32	**44.16** / **87.75**
Sino-Tibetan (3)	X⇒Eng	9.35 / 58.60	9.32 / 65.32	16.59 / 72.34	18.35 / 74.45	16.88 / 74.20	21.36 / 78.52	27.74 / 84.48	11.09 / 71.35	30.88 / 86.50	**35.68** / **87.66**
	Eng⇒X	10.14 / 74.16	2.57 / 54.73	10.74 / 66.74	12.24 / 65.99	9.06 / 65.07	19.92 / 76.04	22.81 / 81.11	10.42 / 73.82	16.85 / 80.74	**32.40** / **88.52**
Other (14)	X⇒Eng	9.71 / 60.43	10.10 / 60.78	5.37 / 47.38	16.00 / 71.15	14.25 / 70.35	25.59 / 82.48	32.62 / 86.21	25.53 / 81.53	35.06 / 86.86	**36.95** / **87.93**
	Eng⇒X	8.42 / 51.57	3.82 / 46.85	1.73 / 29.73	8.19 / 53.20	7.14 / 52.12	20.26 / 74.31	24.04 / 79.59	23.29 / 77.80	28.54 / 85.84	**34.34** / **87.82**

Table 1: Average translation performance of LLMs on different language families. The number in the bracket indicates the number of evaluated languages in the specific language family. Bold text denotes the highest BLEU or COMET score across models. Underlined text denotes the highest BLEU or COMET score across LLMs.

3 Experiment Setup

Dataset We benchmark multilingual translation on FLORES-101 (Goyal et al., 2022) dataset[1], which enables an assessment of model quality on a wide range of languages.

LLMs We evaluate translation performance of eight popular LLMs: XGLM-7.5B (Lin et al., 2022), OPT-175B (Zhang et al., 2022), BLOOMZ-7.1B (Scao et al., 2022), Falcon-7B (Almazrouei et al., 2023), LLaMA2-7B (Touvron et al., 2023), LLaMA2-7B-chat (Touvron et al., 2023), ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023).

ICL strategy For each model, we report its translation performance with eight randomly-picked translation pairs from the corresponding development set as in-context exemplars and "<X>=<Y>" as in-context template. "<X>" and "<Y>" are the placeholder for the source and target sentence. We use line-break as the concatenation symbol. According to our experiment analysis, this ICL strategy serves as a simple but strong recipe. All implementation is based on *OpenICL*[2] (Wu et al., 2023).

Supervised baselines We report the performance of the supervised model M2M-100-12B (Fan et al., 2021) and NLLB-1.3B (Costa-jussà et al., 2022) (distillation version), which are widely-used many-to-many MMT models. We also report the performance of the powerful commercial translation system, Google Translate[3].

Metric Following Goyal et al. (2022), we use SentencePiece BLEU[4] (spBLEU) as evaluation metric, which enables an evaluation of all languages. In addition, we also consider emerging metrics, COMET[5] (Rei et al., 2020) and SEScore[6] (Xu et al., 2022b), which have been shown to correlate well with human judgements.

4 Benchmarking LLMs for Massively Multilingual Machine Translation

In this section, we report results on multilingual machine translation and introduce our main findings about LLMs' translation ability.

The multilingual translation capabilities of LLMs are continually involving Table 1 presents evaluation results[7] grouped by language

[1]We evaluate LLMs on the first 100 sentences of each direction's test set in benchmarking experiment, considering the prohibitive API cost of evaluating massive languages. In analysis experiment, we use full test set.

[2]https://github.com/Shark-NLP/OpenICL

[3]https://translate.google.com/

[4]https://github.com/mjpost/sacrebleu

[5]We compute the score with *wmt22-comet-da* model.

[6]We compute the score with *SEScore-2* (Xu et al., 2022a).

[7]Evaluating with SEScore leads to similar findings, thus we report those results in Appendix A. Detailed results for

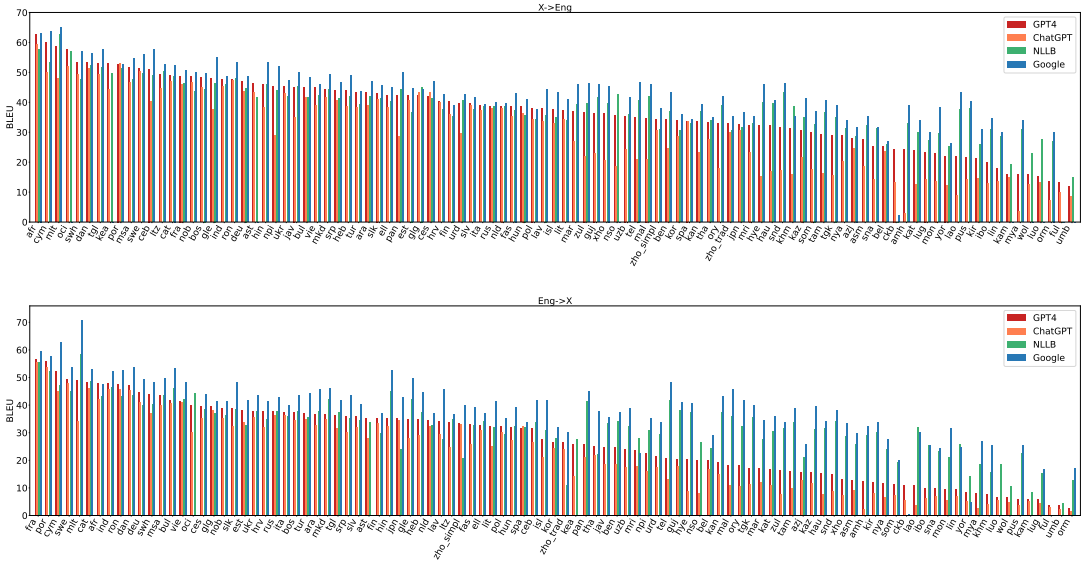

Figure 2: Translation performance (BLEU) of GPT-4, ChatGPT, NLLB and Google Translate on our evaluated languages. "X->Eng" and "Eng->X" denote translating to English and translating from English respectively. In each subfigure, languages are sorted according to BLEU scores of GPT-4.

family. Monolingual pre-trained LLMs present impressive multilingual translation ability, indicating the possibility of aligning multiple languages even with unsupervised data (Garcia et al., 2023). More encouragingly, the multilingual translation capabilities of LLMs are continually improving. The most recent LLMs are reaching new performance heights; for example, LLaMA2-7B outperforms previously released open-source LLMs, and GPT-4 surpasses ChatGPT. Overall, GPT-4 is the best translator among evaluated LLMs and it achieves the highest average BLEU and COMET score on most directions.

Language Family	X⇒Eng	X⇒Fra	X⇒Zho	Eng⇒X	Fra⇒X	Zho⇒X
Indo-Euro-Germanic (8)	48.51	44.23	27.97	40.64	32.34	24.13
Indo-Euro-Romance (8)	47.29	45.16	27.31	44.47	36.05	27.12
Indo-Euro-Slavic (12)	41.19	40.32	25.67	36.06	30.88	23.33
Indo-Euro-Indo-Aryan (10)	37.30	32.81	21.81	21.35	17.26	13.55
Indo-Euro-Other (11)	37.29	35.36	22.70	28.45	22.57	17.50
Austronesian (6)	46.81	39.98	24.40	34.66	25.64	19.52
Atlantic-Congo (14)	28.27	25.02	15.72	13.70	10.42	7.60
Afro-Asiatic (6)	30.48	27.00	17.81	19.36	14.43	10.53
Turkic (5)	31.73	30.90	19.96	20.96	17.80	14.02
Dravidian (4)	33.10	30.61	20.63	18.60	14.47	11.37
Sino-Tibetan (3)	27.74	27.93	20.88	22.81	19.21	16.30
Other (14)	32.62	31.26	21.25	24.04	20.03	16.37

Table 2: Translation performance (BLEU) of GPT-4 on English-centric, French-centric and Chinese-centric translation.

each translation direction are listed in Appendix B.

LLM's capability is unbalanced across languages In Table 1, we observe a similar trend for all evaluated LLMs: they perform better at translating into English than translating into non-English. LLM's capability on non-English languages is also unbalanced. For languages that are similar to English, e.g, Indo-European-Germanic languages, LLMs achieve impressive results. For languages that are dissimilar to English, e.g., Sino-Tibetan languages, LLMs often produce less decent results.

Table 2 presents another clue, where we evaluate GPT-4 on French-centric and Chinese-centric translation. Compared to English-centric translation, GPT-4 faces greater challenge when it comes to non-English-centric translation, which again indicates LLM's unbalanced translation ability across languages.

LLMs still lag behind the strong supervised baseline, especially on low-resource languages Figure 2 shows the translation performance of the supervised systems and GPT-4 on each language. In 40.91% translation directions, GPT-4 has achieved higher BLEU scores than NLLB, indicating the promising future of this new translation paradigm. But on long-tail low-resource languages, GPT-4 still lags behind NLLB, let alone Google Translate.

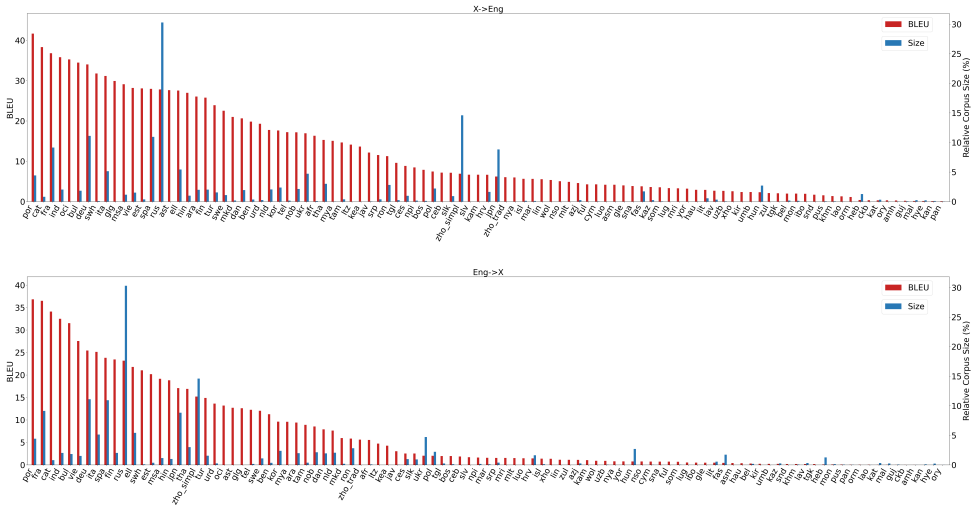

Figure 3: Translation performance (BLEU) of XGLM on evaluated languages and the corpus size of each language relative to English pre-training corpus. In each subfigure, languages are sorted according to BLEU scores of XGLM.

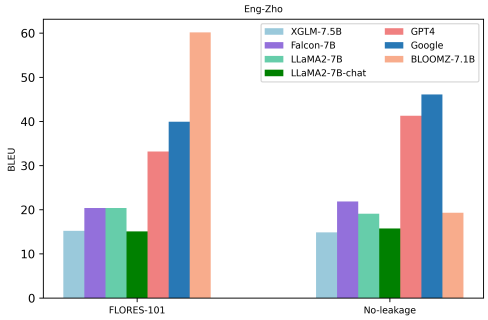

Figure 4: Translation performance of different models on FLORES-101 test set and our annotated no-leakage evaluation set NEWS2023.

Data leakage issue should be considered before evaluating LLMs on public datasets. We do not include BLOOMZ's performance on FLORES-101 in our report because BLOOMZ is instruction-tuned with XP3 dataset (Scao et al., 2022), which includes FLORES-200 dataset. Thus BLOOMZ may have been exposed to test cases from FLORES-101 during training. If so, the evaluation results can not precisely reflect its translation ability (Elangovan et al., 2021).

To illustrate this concern, we take 1000 English sentences from the most recent news spanning August 2023 to October 2023[8], and ask human experts to translate them into Chinese and construct a bilingual no-leakage evaluation set, named NEWS2023. Figure 4 shows that BLOOMZ's performance significantly deteriorates on this no leakage set, whereas other models maintain a consistent performance across both datasets. Through this example, we wish to draw the community's attention to the potential data leakage issue when evaluating large language models.

5 Analyzing Factors That Influence LLM's Translation Performance

To better understand how LLM acquires translation ability and which factors have influence on its performance, we conduct in-depth analysis. For analysis, we choose XGLM-7.5B as an example[9]. Note that, when studying a certain factor, we keep the remaining factors unchanged.

5.1 Findings on Pre-training Corpus Size

LLM can acquire translation ability in a resource-efficient way. As XGLM authors report data distribution of their pre-training corpus, we can investigate the relationship between translation performance and corpus size (Figure 3). We find that for low-resource languages, e.g., Catalan (cat) and Swahili (swh), XGLM can generate moderate translation, showing that LLM can build bilingual mapping between non-English and English with a few non-English monolingual resources (less than 1% of English resources). Even on unseen

[8]The news were collected from BBC news, Fox news, ABC news and Yahoo news.

[9]We choose XGLM for three reasons: (1) XGLM has a multilingual focus and covers many languages, which can be seen as a representative of multilingual LLM. (2) XGLM-7.5B is an open-source medium-sized LLM. It is more affordable to run experiments with it than large-sized LLM or close-source LLM. (3) The composition of the XGLM's pre-training corpus is clear, allowing us to analyze the relationship between translation ability and corpus size.

In-context Template	Deu-Eng	Eng-Deu	Rus-Eng	Eng-Rus	Rus-Deu	Deu-Rus	Average
`reasonable instructions:`							
<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 instructions:`							
<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

Table 3: Translation performance (BLEU) of using different templates for in-context learning. The number of in-context exemplars is fixed at eight in this experiment. "<X>" and "<Y>" denote the placeholder for source and target sentence respectively. "[SRC]" and "[TGT]" represent the placeholder for source and target language name in English. Bold text denotes the highest score along the column.

languages, e.g., Occitan (oci) and Asturian (ast), XGLM can translate through ICL. These observations indicate a potential advantage of the novel translation paradigm: LLM can learn to translate in a resource-efficient way.

5.2 Findings on In-context Template

The good performance of LLMs relies on carefully-designed template The initial step of applying in-context learning for translation is determining the template. We find that the translation performance varies greatly with different templates (Table 3), where the largest gap in the average performance is up to 16 BLEU. The best template for each direction is also different. Among these templates, "<X>=<Y>" achieves the highest average BLEU score. "[SRC]: <X> \n [TGT]: <Y>" achieves the lowest score, although it is a commonly-used template for prompting other LLMs, e.g., PaLM (Vilar et al., 2022), GLM (Zhang et al., 2023). Such phenomena indicate that the template plays a vital role in ICL and it may be challenging to design a universally optimal template for different LLMs and translation directions.

Even unreasonable template can instruct LLM to generate decent translation A common intuition of ICL is that the template instructs LLMs to do the target task (Brown et al., 2020), e.g., the template "<X> can be translated to <Y>" instructs the LLM to perform translation task. However, we find that wrapping translation exemplars with task-unrelated template can also serve as an effective prompt. For example, the template like "<X> can be summarized as <Y>" can also instruct LLM to generate translation, rather than guiding it to generate summarization. Given the fact that these unreasonable template are also effective, the community may not fully understand the role of in-context-template.

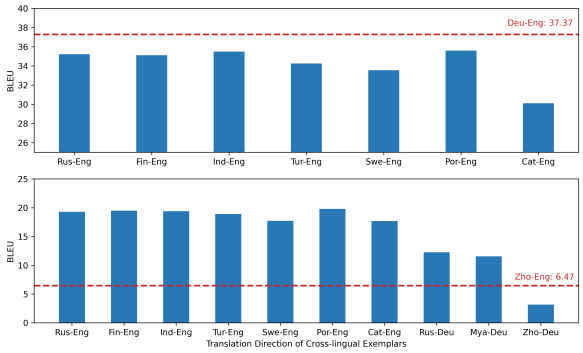

Figure 5: Effects of using cross-lingual exemplars.

5.3 Findings on In-context Exemplar

Cross-lingual exemplars help for certain translation directions Translation direction of the exemplar is a unique factor in machine translation. We find that using cross-lingual exemplars does not always causes worse performance and show two cases in Figure 5. When using cross-lingual exemplars for German-English translation, the translation performance degenerates. But when using cross-lingual exemplars for low-resource Chinese-English translation (illustrated in Appendix D), XGLM's translation performance usually improves

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

Table 4: Translation performance (BLEU) of XGLM when using different contents as in-context exemplars. "Consistency" column denotes whether source and target sentence are semantically consistent. "Granularity" column denotes whether the exemplar is a sentence-level pair. "Diversity" column denotes whether exemplars in the context are different from each other.

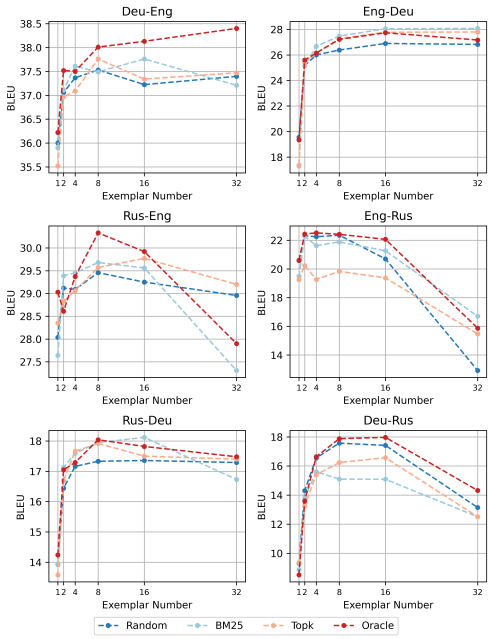

Figure 6: Effects of selecting varying number of in-context exemplars according to different strategies.

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

Table 5: Effects of reversing in-context examples' translation direction. "Rev ratio" means the number of exemplars that are reversed. "Head" and "Tail" represents reversing the exemplars in the head and tail of the prompt respectively.

significantly, even when both source and target language is changed. This phenomenon indicates the potential usage of cross-lingual exemplars in a broader range of tasks (Lin et al., 2022), and we will explore more about this in the future.

Semantically-related exemplars does not brings more benefits than randomly-picked exemplars In this paper, we use development set for exemplar selection, which has been found to be a high-quality candidate pool (Vilar et al., 2022), and we compare four ways of selecting in-context exemplars, namely *Random*[10], *BM25*[11], *TopK*[12] and *Oracle*[13].

[10] *Random*: picking exemplars on a random basis.

[11] *BM25*: selecting exemplars whose source sentences are similar to the test case's source sentence according to BM25.

[12] *TopK*: selecting exemplars whose source sentences are similar to the test case's source sentence according to the similarity of sentence embedding.

[13] *Oracle*: selecting exemplars whose target sentences are similar to the test case's according to sentence embedding, which can be seen as the upper bound of selection strategy.

Effects of selecting varying number of in-context exemplars with different approaches are shown in Figure 6. The general trend in all dataset is similar. As the number of examples grows from 1 to 8, the BLEU score increases rapidly. Afterwards, the translation performance plateaus regardless of selection strategy. When more exemplars are added, e.g., 32 exemplars, the BLEU score usually starts to decline, shows an opposite phenomenon against the observation in natural language understanding tasks (Li et al., 2023).

Compared to semantically-related exemplars, randomly-picked exemplars gives comparable translation performance. Even the performance of oracle selection is on par with random selection. Based on these observations, we suggest that translation exemplars can teach LLM to translate but LLM may struggle to acquire helpful translation knowledge from semantically-related exemplars.

Exemplars teach LLM the core feature of translation task To better understand how ICL exemplars influence LLM to understand the translation task, we observe LLM's translation behaviour under abnormal in-context exemplars (Table 4).

We can see that LLM completely fails when mismatched translation is used as exemplars, indicating that LLM needs to learn from the context to keep source and target sentence semantically consistent. Word-level[14] and document-level[15] translation exemplar degenerates LLM's translation performance, which demonstrates that the translation granularity of exemplar matters as well. Another interesting phenomenon is that LLM performs worse when duplicated translation is used as the exemplar, indicating that keeping in-context exemplars diverse is also important. In general, these comparison results show that LLM learns the core feature of translation task through in-context learning.

The exemplar in the tail of the prompt has more impact on the LLM's behaviour During our analysis, we find that reversing the translation direction of exemplars will cause LLM to fail. Based on this observation, we conduct experiments to investigate the importance of different parts of the prompt (Table 5). We find that reversing exemplars in the tail of the prompt consistently produced worse results compared to reversing exemplars in the head, which suggests that exemplars in the tail of the prompt have larger influence on LLM's behavior.

6 Related Work

In-context learning for machine translation Using LLMs for multilingual machine translation is attracting more and more attention. Lin et al. (2022) evaluate GPT-3 and XGLM-7.5B on 182 directions. Bawden and Yvon (2023) evaluates BLOOM on 30 directions. Bang et al. (2023), Jiao et al. (2023) and Hendy et al. (2023) evaluate ChatGPT on 6 to 18 directions. In this paper, we thoroughly evaluate multilingual translation performance of popular LLMs on 102 languages and 606 directions and compare them with state-of-the-art translation engines, such as NLLB and Google Translate, which provides a more comprehensive benchmark result and highlights the challenges involved in optimizing this emerging translation paradigm.

To find better ICL recipe for machine translation, many efforts have been put into designing exemplars selection strategy (Agrawal et al., 2022; Zhang et al., 2023; Moslem et al., 2023). Similar to the findings of Zhang et al. (2023), we find that random selection is a simple but effective strategy. We also find that even oracle selection can not result in consistently better performance. Wei et al. (2022a) shows few-shot exemplars improve translation performance. And we further demonstrate the dynamic variations of translation performance with the number of in-context exemplars and the usage of cross-lingual exemplars. Besides, Vilar et al. (2022) find that using a high-quality pool, e.g., development set, for ICL example selection is better and Zhang et al. (2023) analyze why the quality of translation exemplars matters. In this paper, we reveal how in-context exemplars teach LLM to translate by analyzing LLM's behaviour under different kinds of exemplars.

Multilingual machine translation Developing a bilingual translation system for each direction becomes impossible when the number of supporting languages increases. Therefore, multilingual machine translation is proposed (Johnson et al., 2017). But how to build a high-quality yet efficient MMT system remains an on-going challenge (Costa-jussà et al., 2022; Yuan et al., 2023; Guerreiro et al., 2023). In this paper, we focus on LLM and reveal its potential in MMT.

7 Conclusion

In this paper, we evaluate the multilingual translation ability of popular LLMs, including ChatGPT and GPT-4, on 102 languages and 606 directions, which presents the advantages and challenges of LLMs for MMT. We find that translation capabilities of LLMs are continually involving and GPT-4 reaches new performance height. However, even for GPT-4, it still face challenge on low-resource languages. In our analysis, we find that LLMs exhibit new working patterns when used for MMT. For example, instruction semantics can be ignored during in-context learning and cross-lingual exemplars can provide better task instruction for low-resource translation. More importantly, we find that LLM can acquire translation ability in a resource-efficient way, which indicates the promising future of LLM in multilingual machine translation.

Limitations

In this paper, we mainly evaluate LLM's English-centric, French-centric and Chinese-centric translation ability. In the future, we would like to investigate more translation directions, e.g., Russian-centric translation, Arabic-centric trans-

[14]We select word pairs from open-source *fasttext* dictionary.
[15]We select document translation from Europarl dataset.

lation, which could bring more findings concerning with LLM's translation ability.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023. Falcon-40b: an open large language model with state-of-the-art performance, 2023. *URL https://huggingface.co/tiiuae/falcon-40b*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems (NeurIPS)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research (JMLR)*.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *arXiv preprint arXiv:2302.01398*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics (TACL)*.

Nuno M Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *arXiv preprint arXiv:2303.16104*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems (NeurIPS)*.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics (TACL)*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations (ICLR)*.

Mukai Li, Shansan Gong, Jiangtao Feng, Yiheng Xu, Jun Zhang, Zhiyong Wu, and Lingpeng Kong. 2023. In-context learning with many demonstration examples. *arXiv preprint arXiv:2302.04931*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian hLi. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. *Interspeech*.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

OpenAI. 2022. https://openai.com/blog/chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, Alexander Podolskiy, Grigory Arshinov, Andrey Bout, Irina Piontkovskaya, Jiansheng Wei, Xin Jiang, Teng Su, Qun Liu, and Jun Yao. 2023. Pangu-$sigma$: Towards trillion parameter language model with sparse heterogeneous computing. *arXiv preprint arXiv:2303.10845*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846.

Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. 2023. Openicl: An open-source framework for in-context learning. *arXiv preprint arXiv:2303.02913*.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2022a. Sescore2: Retrieval augmented pretraining for text generation evaluation. *arXiv preprint arXiv:2212.09305*.

Wenda Xu, Yi-Lin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022b. Not all errors are equal: Learning text generation metrics using stratified error synthesis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-mt: Towards detachable models in massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Evaluating LLM's translation performance with SEScore

Table 6 presents average SEScore of LLMs on different language families. Currently, SEScore mainly supports evaluating English translation. Thus we evaluate LLM's performance on translating other languages to English.

B Detailed Results on Each Language

We report detailed results of our evaluated models in Table 7 (BLEU), Table 8 (COMET), Table 9 (SEScore) and Figure 8. One thing that needs to be mentioned is that BLEU supports all translation directions, whereas COMET and SEScore only support a subset of these translation directions.

C Lists of Language

We evaluate 102 languages in this paper. Table 10 lists the name, ISO code and language family of these languages.

D Cross-lingual Exemplars

In Figure 7, we show an example of using cross-lingual in-context exemplars (Russian-English exemplars for Chinese-English translation).

[Input]

Этот фильм с участием Райана Гослинга и Эммы Стоун получил номинации во всех главных категориях.=The movie, featuring Ryan Gosling and Emma Stone, received nominations in all major categories.

"Теперь у нас есть четырёхмесячные мыши, у которых больше нет диабета", — добавил он.="We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

Гослинг и Стоун получили номинации на лучшего актера и актрису соответственно.=Gosling and Stone received nominations for Best Actor and Actress respectively.

Находка также позволяет ознакомиться с эволюцией перьев у птиц.=The find also grants insight into the evolution of feathers in birds.

Канцелярия губернатора сообщила, что 19 из раненных были офицерами полиции.=The governor's office said nineteen of the injured were police officers.

Стандарт 802.11n работает на обоих частотах – 2.4 ГГц и 5.0 ГГц.=The 802.11n standard operates on both the 2.4Ghz and 5.0Ghz frequencies.

Он сказал, что создал дверной звонок, работающий от WiFi.=He built a WiFi door bell, he said.

В конце 2017 года Симинофф появился на торговом телеканале QVC.=In late 2017, Siminoff appeared on shopping television channel QVC.

伊拉克研究小组于格林尼治时间 (GMT) 今天 12 点提交了报告。=

[Output]

The Iraqi research team submitted a report at Greenwich time (GMT) today at 12 noon.

Figure 7: An example of using cross-lingual in-context exemplars

E Used Scientific Artifacts

Below lists scientific artifacts that are used in our work. For the sake of ethic, our use of these artifacts is consistent with their intended use.

- *OpenICL (Apache-2.0 license)*, a framework that provides an easy interface for in-context learning.
- *Transformers (Apache-2.0 license)*, a framework that provides thousands of pretrained models to perform tasks on different modalities such as text, vision, and audio.

Language Family	Direction	Translation Performance (SEScore)								
		XGLM-7.5B	OPT-175B	Falcon-7B	LLaMA-7B	LLaMA-7B-Chat	ChatGPT	GPT4	M2M-12B	NLLB-1.3B	Google
Indo-Euro-Germanic (8)	X⇒Eng	-11.78	-6.00	-8.34	-5.41	-5.90	-2.52	-2.16	-3.15	-2.78	**-1.85**
Indo-Euro-Romance (8)	X⇒Eng	-6.54	-4.01	-5.57	-3.72	-4.14	-2.30	**-2.08**	-3.08	-2.54	-2.12
Indo-Euro-Slavic (12)	X⇒Eng	-14.29	-10.31	-13.46	-5.11	-5.75	-3.55	-3.17	-4.21	-3.70	**-2.80**
Indo-Euro-Indo-Aryan (10)	X⇒Eng	-16.45	-22.15	-21.65	-17.15	-19.46	-7.64	-4.69	-11.77	-3.53	**-2.80**
Indo-Euro-Other (11)	X⇒Eng	-18.36	-17.81	-18.09	-13.61	-15.42	-6.74	-4.62	-7.57	**-3.75**	-4.40
Austronesian (6)	X⇒Eng	-14.06	-10.08	-12.30	-9.61	-10.48	-4.48	-3.03	-5.37	-3.47	**-2.56**
Atlantic-Congo (14)	X⇒Eng	-19.42	-17.61	-18.44	-17.59	-18.48	-12.38	-9.34	-14.16	-6.88	**-5.75**
Afro-Asiatic (6)	X⇒Eng	-18.85	-18.91	-19.17	-16.61	-17.66	-12.16	-8.28	-14.41	-4.46	**-3.49**
Turkic (5)	X⇒Eng	-17.15	-16.99	-18.66	-15.50	-16.47	-7.63	-5.50	-15.29	-4.89	**-3.93**
Dravidian (4)	X⇒Eng	-16.52	-22.58	-21.91	-20.18	-21.96	-9.26	-5.35	-13.69	-3.76	**-3.07**
Sino-Tibetan (3)	X⇒Eng	-19.41	-15.20	-12.37	-11.33	-12.01	-10.43	-6.79	-11.93	-5.50	**-4.30**
Other (14)	X⇒Eng	-16.74	-16.56	-18.70	-13.05	-14.17	-8.51	-6.07	-6.91	-4.94	**-3.80**

Table 6: Average SEScore of LLMs on different language families. The number in the bracket indicates the number of evaluated languages in the specific language family. Bold text denotes the highest SEScore across models. Underlined text denotes the highest SEScore across LLMs.

Language Family	Language	X⇒Eng (BLEU)										Eng⇒X (BLEU)									
		XGLM-7.5B	OPT-175B	Falcon-7B	LLaMA2-7B	LLaMA2-7B-Chat	ChatGPT	GPT4	M2M-12B	NLLB-1.3B	Google	XGLM-7.5B	OPT-175B	Falcon-7B	LLaMA2-7B	LLaMA2-7B-Chat	ChatGPT	GPT4	M2M-12B	NLLB-1.3B	Google
Indo-European-Germanic (8)	afr	16.34	48.49	34.73	47.89	42.89	59.28	62.65	52.86	57.76	63.15	5.56	20.75	14.45	22.98	20.42	42.18	48.02	41.41	43.39	47.83
	dan	20.65	43.54	35.31	48.33	45.83	51.23	53.18	48.32	52.35	56.44	7.91	26.81	14.80	32.79	28.19	45.49	47.46	45.12	43.81	53.99
	nld	17.78	31.25	26.87	34.46	33.03	38.10	38.60	34.52	38.68	39.66	7.64	21.38	16.69	24.89	20.80	32.57	34.66	31.79	32.93	37.05
	deu	34.03	39.15	34.60	41.94	39.44	43.56	47.04	42.79	44.79	48.52	25.44	23.38	20.65	30.46	26.01	41.02	44.69	40.18	40.20	49.32
	isl	5.65	12.68	8.18	15.41	12.28	32.98	37.58	29.47	35.07	43.19	1.40	3.10	2.77	5.13	5.53	21.26	27.89	27.80	31.04	41.80
	ltz	14.13	17.96	13.60	21.87	18.36	44.57	49.20	40.04	50.37	52.52	4.74	5.54	5.10	6.32	5.72	24.65	33.89	28.04	35.08	36.80
	nob	17.19	39.45	28.38	41.91	42.08	46.62	48.51	45.38	43.76	49.94	8.55	23.18	12.90	26.01	20.35	35.44	39.10	37.09	36.33	41.40
	swe	22.54	44.67	37.30	46.47	44.62	50.32	51.34	48.37	49.50	55.86	12.04	27.00	18.12	33.69	28.49	48.09	49.39	47.02	45.00	53.96
Average		18.54	34.65	27.37	37.28	34.82	45.83	48.51	42.72	46.54	51.16	9.16	18.89	13.19	22.78	19.44	36.34	40.64	37.30	38.47	45.27
Indo-European-Romance (8)	ast	27.65	32.20	28.84	33.88	30.90	43.18	46.41	39.06	41.65	-1.00	12.70	13.11	10.96	12.89	11.24	28.24	35.45	33.43	34.01	-1.00
	cat	38.33	41.45	27.52	44.48	40.97	47.04	49.10	44.21	48.72	52.46	34.10	23.49	13.95	36.18	35.31	46.33	48.34	48.49	48.79	53.23
	fra	36.81	43.02	41.62	44.11	41.15	46.13	48.81	43.99	46.23	50.68	36.49	37.97	43.87	42.86	39.60	55.71	56.80	53.59	55.73	59.73
	glg	29.93	36.57	29.30	37.98	35.43	43.33	42.18	38.13	45.12	44.18	12.60	18.53	12.30	16.07	14.38	38.07	39.54	38.29	37.11	41.49
	oci	35.27	41.41	36.11	42.89	37.45	51.86	57.73	48.03	56.93	-1.00	13.20	8.90	7.60	12.76	11.62	30.33	40.20	39.40	44.45	-1.00
	por	41.67	44.64	44.49	48.14	45.47	53.09	52.81	48.76	51.20	52.68	36.83	37.72	34.62	42.85	38.70	53.95	55.89	53.75	52.29	57.85
	ron	11.27	41.33	34.49	44.24	40.83	47.31	47.53	45.87	47.85	53.18	5.85	31.35	14.97	33.08	28.31	45.87	47.62	47.99	43.42	52.76
	spa	27.98	30.81	30.13	33.09	30.51	33.48	33.76	30.63	32.91	34.36	23.82	23.35	21.93	25.83	24.84	32.31	31.88	28.93	32.08	33.76
Average		24.83	36.79	30.72	39.19	36.33	45.76	47.90	42.53	46.43	43.43	15.55	21.60	16.61	25.30	22.47	38.84	42.55	40.14	40.98	41.19
Indo-European-Slavic (12)	bel	1.98	4.48	1.88	12.85	9.48	23.71	25.12	15.62	26.00	27.03	0.31	0.35	0.39	3.39	1.89	16.95	20.13	13.59	24.55	29.34
	bos	7.88	34.37	21.26	39.24	37.13	44.86	48.34	41.24	44.47	49.75	1.97	18.05	7.41	23.37	18.71	34.44	37.52	33.78	37.77	43.67
	bul	34.48	11.48	8.07	38.18	34.32	41.65	44.97	40.50	41.60	48.32	31.53	2.83	3.11	26.38	20.13	40.78	42.02	49.44	46.38	53.32
	hrv	6.66	33.37	19.48	36.35	34.68	40.02	40.42	36.28	37.62	42.60	1.44	15.71	6.19	21.96	17.66	31.90	37.84	32.54	34.94	41.63
	ces	8.84	32.26	22.03	39.44	35.74	43.25	42.08	41.87	41.42	47.00	2.54	15.47	8.09	27.30	21.73	35.22	39.72	37.21	38.62	44.11
	mkd	21.00	8.32	5.63	33.36	27.81	41.76	44.36	39.59	44.34	49.21	5.97	1.52	2.06	12.80	8.58	34.94	36.69	42.38	42.31	46.31
	pol	7.46	28.63	23.95	33.02	31.44	34.31	38.12	32.65	34.27	37.74	2.02	14.15	7.96	20.79	17.93	30.16	32.27	29.26	29.67	35.29
	rus	27.83	18.80	14.26	33.44	31.92	38.04	38.75	32.73	38.60	40.09	23.18	6.48	3.49	25.54	21.50	36.45	37.71	39.69	37.86	43.10
	srp	11.56	6.57	4.70	36.97	33.34	40.71	44.09	37.56	41.40	46.75	1.55	0.86	1.30	24.58	19.85	30.39	36.18	30.00	35.35	43.56
	slk	7.15	30.21	16.86	31.50	29.03	40.92	43.13	38.57	41.28	45.71	2.54	10.24	5.80	13.66	10.30	32.48	38.78	37.84	38.73	48.36
	slv	6.67	25.64	13.08	33.26	29.52	39.04	39.70	35.88	37.73	41.69	1.71	9.10	4.78	17.98	16.37	32.04	36.03	36.89	34.77	40.58
	ukr	16.95	15.80	6.63	40.37	36.89	42.95	45.16	37.89	41.97	47.44	2.04	3.38	1.49	25.17	19.08	35.53	37.87	37.54	37.80	43.74
Average		19.84	29.95	23.19	36.97	34.02	42.97	45.02	39.67	43.34	43.51	11.63	15.85	11.35	23.13	19.76	36.17	39.77	37.94	39.09	41.86
Indo-European-Indo-Aryan (10)	asm	4.18	1.11	1.17	3.82	1.27	18.58	27.47	-1.00	32.32	35.35	0.42	0.05	0.05	0.21	0.07	9.08	12.74	-1.00	26.02	29.77
	ben	19.84	1.12	1.66	6.72	2.71	24.63	34.23	30.60	36.97	43.37	11.27	0.03	0.11	2.09	0.78	18.65	24.74	28.39	34.31	37.66
	guj	0.21	1.06	1.65	1.49	1.61	22.78	36.44	0.90	41.76	45.97	0.03	0.02	0.04	0.21	0.11	18.05	20.65	7.32	38.37	40.99
	hin	26.99	1.17	1.26	21.04	14.89	38.15	45.88	40.72	45.83	53.17	18.81	0.42	0.27	5.84	5.18	32.44	35.30	40.54	44.97	52.86
	mar	5.63	0.87	1.00	7.37	4.78	26.94	37.08	27.29	39.25	46.02	1.58	0.06	0.07	2.17	1.83	12.22	17.13	18.27	27.66	34.71
	npi	8.47	2.31	3.17	9.88	6.62	28.83	45.25	19.00	44.01	51.91	1.63	0.12	0.14	2.14	1.65	16.16	22.73	4.08	30.96	35.39
	ory	0.31	0.82	1.14	1.35	1.33	17.83	33.07	0.64	39.02	42.00	0.01	0.06	0.02	0.05	0.02	10.70	18.12	0.60	32.57	41.71
	pan	0.13	1.09	1.17	2.09	1.46	28.65	42.28	24.92	44.34	49.86	0.06	0.06	0.01	0.21	0.17	21.38	25.73	14.85	41.57	45.16
	snd	1.70	1.72	0.65	4.27	3.25	17.29	31.53	8.31	43.32	46.23	0.20	0.39	0.31	0.82	0.60	8.75	14.97	13.15	34.34	38.15
	urd	19.31	0.74	1.09	8.76	4.95	29.53	39.72	23.94	40.67	42.69	13.63	0.20	0.29	2.37	2.03	17.58	21.43	18.17	29.65	34.04
Average		16.91	22.38	17.45	29.00	26.20	38.33	42.99	33.85	42.66	44.07	9.82	11.71	8.40	17.47	14.89	30.99	34.92	31.76	37.76	41.12
Indo-European-Other (11)	hye	0.15	0.32	0.74	3.83	2.05	15.30	32.20	20.70	39.99	45.84	0.02	0.05	0.01	1.19	1.53	9.02	20.47	9.89	37.54	40.91
	ell	27.54	9.42	5.70	24.18	17.56	38.39	42.36	35.74	40.41	44.84	21.79	1.07	0.51	2.88	2.37	31.12	32.90	36.02	34.35	37.27
	gle	4.02	10.49	8.63	17.98	13.61	37.74	47.94	3.24	46.48	54.95	0.50	1.46	2.18	4.34	4.72	28.01	34.93	0.23	42.37	49.89
	cym	4.27	10.74	8.46	18.99	12.89	49.92	60.07	29.28	53.33	63.77	0.74	2.66	3.37	5.31	5.20	44.97	52.37	21.91	47.44	63.00
	ita	31.17	32.71	33.41	36.30	35.60	37.32	38.85	34.85	38.69	39.15	25.14	23.95	25.79	27.18	26.06	36.39	37.66	34.86	36.01	40.12
	lav	2.69	7.00	4.73	13.27	8.75	33.54	37.92	34.06	35.79	44.38	0.19	1.76	1.76	2.92	2.24	29.39	34.34	35.58	27.75	46.01
	lit	2.90	7.97	7.60	12.66	11.60	34.34	37.41	33.45	33.80	41.07	0.50	2.08	2.24	4.35	3.48	25.20	32.60	36.08	32.23	41.55
	pus	1.56	1.82	3.05	5.03	4.78	14.30	21.46	24.52	37.97	40.35	0.09	0.20	0.18	0.80	1.16	3.92	6.13	14.14	22.66	25.58
	fas	3.79	2.01	2.58	16.97	12.42	35.30	38.60	32.29	37.16	43.12	0.45	0.12	0.50	3.90	3.70	25.92	32.98	30.11	32.92	39.16
	ckb	0.34	1.48	0.84	2.94	2.34	13.39	24.40	-1.00	-1.00	2.17	0.03	0.11	0.05	0.73	1.07	5.64	11.19	-1.00	-1.00	0.59
	tgk	2.06	1.83	1.65	4.84	4.45	15.41	29.01	-1.00	35.09	38.88	0.18	0.63	0.63	1.39	1.57	11.33	17.37	-1.00	35.83	39.89
Average		14.75	19.11	15.12	25.69	22.89	36.36	41.71	31.27	41.20	43.54	8.63	9.78	7.27	14.67	12.63	29.16	33.47	29.05	36.39	40.54
Austronesian (6)	ceb	7.18	29.10	16.81	23.15	20.83	40.33	51.12	32.93	48.93	57.74	1.86	8.63	6.63	9.49	9.68	26.81	31.65	24.07	33.96	41.87
	tgl	9.61	35.32	22.90	32.40	28.09	49.30	53.09	36.16	51.78	57.79	1.97	15.27	9.80	14.25	12.39	31.58	36.43	27.83	37.46	41.83
	ind	35.82	33.73	27.85	41.10	38.97	45.33	47.54	43.08	46.10	48.65	32.49	20.28	14.82	30.36	26.12	45.80	47.97	43.89	46.40	52.34
	jav	12.17	12.69	9.39	13.80	13.61	34.84	45.14	34.50	45.21	50.08	3.04	3.58	4.22	7.89	7.41	34.62	24.78	26.07	33.54	35.80
	msa	29.11	33.27	28.05	37.03	35.28	46.52	51.61	45.37	47.62	54.68	19.15	14.40	12.62	21.17	17.87	40.13	43.49	41.31	43.61	49.89
	mri	3.29	9.48	6.71	12.73	9.54	23.39	32.34	-1.00	32.84	35.13	1.54	1.92	3.26	4.39	6.26	18.06	23.67	-1.00	28.05	22.69
Average		14.91	19.82	15.50	25.80	23.05	36.75	42.27	31.33	41.66	44.31	8.78	9.88	7.41	14.66	12.70	29.27	33.60	28.83	36.47	40.56
Atlantic-Congo (14)	lug	3.33	8.12	6.18	7.52	7.75	14.11	23.40	7.19	27.17	29.91	0.53	0.54	1.11	1.77	2.56	4.61	5.94	1.62	15.55	16.82
	sho	1.92	5.21	5.36	7.05	7.33	12.99	19.79	16.28	31.05	34.50	0.51	1.09	2.32	1.82	2.54	6.27	9.99	13.53	25.60	25.47
	kea	13.65	26.18	14.53	21.66	21.07	44.40	53.06	-1.00	49.77	-1.00	4.27	5.94	4.97	6.46	5.38	14.34	25.99	-1.00	27.85	-1.00
	kam	6.66	9.85	7.63	8.40	10.84	14.87	16.02	-1.00	19.23	-1.00	1.05	1.26	1.61	1.85	3.45	5.37	6.07	-1.00	8.58	-1.00
	kin	5.56	8.54	7.11	7.07	8.49	13.51	17.88	8.88	28.61	29.85	1.14	1.36	1.54	1.94	3.36	7.18	9.67	1.14	25.93	24.88
	nso	5.05	8.73	7.92	9.25	7.84	18.61	35.60	11.39	42.65	-1.00	0.76	1.32	1.08	2.35	2.66	8.20	20.14	5.54	26.54	-1.00
	nya	5.98	8.88	7.27	8.05	9.29	20.21	28.84	-1.00	31.37	33.87	0.80	1.60	1.45	2.69	3.45	6.87	11.61	-1.00	23.95	27.64
	sna	3.85	9.05	6.76	8.74	8.69	14.27	25.25	-1.00	31.16	31.69	0.73	1.14	1.48	1.61	3.31	7.09	9.82	-1.00	23.32	24.41
	swh	31.78	11.86	8.19	11.79	9.41	49.29	53.27	42.13	47.58	56.98	21.03	2.27	2.30	3.31	4.39	37.19	44.01	38.05	40.43	48.25
	umb	2.36	4.94	3.68	4.32	4.86	8.44	11.83	-1.00	14.87	-1.00	0.23	0.68	0.69	0.98	1.52	2.32	3.83	-1.00	4.46	-1.00
	wol	5.35	7.92	6.42	8.80	7.64	12.47	15.82	10.16	22.82	-1.00	0.92	1.67	1.78	3.36	3.59	4.95	6.57	1.21	10.73	-1.00
	xho	2.56	7.49	6.06	7.72	8.66	20.69	36.15	26.94	39.66	45.45	1.37	1.37	2.89	2.61	2.59	7.56	13.11	16.61	28.65	33.60
	yor	3.21	6.05	6.15	5.84	7.15	12.35	22.08	6.27	25.39	26.23	0.78	1.05	1.48	2.04	2.52	5.16	8.63	3.82	14.41	4.78
	zul	2.10	5.61	4.43	6.50	7.10	21.89	36.77	23.45	39.49	46.21	1.13	1.10	1.75	1.55	1.90	7.66	16.36	14.85	31.87	33.94
Average		13.24	17.66	13.77	22.34	20.20	33.32	39.43	27.12	39.74	40.10	7.51	8.20	6.29	12.18	10.75	25.14	29.56	24.31	33.53	35.73
Afro-Asiatic (6)	amh	0.29	0.45	0.93	0.94	1.63	2.97	24.14	15.75	32.98	38.99	0.02	0.04	0.02	0.02	0.07	2.22	12.35	12.38	29.12	32.55
	ara	26.06	1.03	1.81	22.35	13.99	38.94	43.29	35.24	42.05	46.87	9.42	0.27	0.27	4.81	3.73	32.64	36.91	31.10	37.81	45.89
	ful	4.28	7.21	6.47	6.69	8.25	10.02	13.33	6.25	-1.00	-1.00	0.72	1.62	1.61	2.61	2.69	3.11	3.89	0.42	-1.00	-1.00
	mlt	4.90	14.75	11.83	21.92	17.68	48.08	58.72	-1.00	62.54	65.03	1.52	3.79	4.33	8.28	7.57	34.42	49.04	-1.00	58.40	70.95
	orm	1.14	2.85	2.47	3.51	3.32	7.32	13.41	-1.00	26.83	30.10	0.05	0.29	0.78	0.95	1.43	1.72	2.71	-1.00	12.69	17.38
	som	3.55	9.30	5.71	7.07	7.06	17.72	29.99	4.76	32.76	36.85	0.70	2.38	1.39	2.68	2.94	7.31	11.25	5.06	19.45	20.23
Average		12.72	16.72	13.06	21.39	19.28	32.32	38.71	25.75	39.18	39.78	7.08	7.65	5.90	11.47	10.14	24.21	28.75	22.99	32.94	35.35
Turkic (5)	azj	4.61	7.01	3.40	8.63	6.56	24.64	27.80	9.33	28.45	31.77	1.12	1.30	1.67	2.24	2.41	12.97	15.79	10.28	21.23	25.92
	kaz	3.62	1.46	1.63	6.55	6.83	21.74	30.65	3.81	34.85	41.16	0.23	0.26	0.48	1.26	1.45	11.92	15.62	13.30	31.42	39.55
	kir	2.37	1.40	1.65	4.83	5.89	14.49	21.31	-1.00	26.00	30.85	0.24	0.27	0.71	2.21	1.74	8.17	12.09	-1.00	30.39	33.87
	tur	23.91	24.39	10.05	21.75	19.93	38.14	43.43	36.76	39.42	43.49	14.90	10.11	4.56	8.82	7.82	35.05	37.05	29.67	35.58	44.29
	uzb	2.66	5.17	4.00	5.77	5.17	24.21	35.45	2.37	35.89	41.63	0.90	0.96	1.31	1.88	2.03	17.54	24.26	2.07	32.25	39.07
Average		12.39	16.17	12.50	20.65	18.63	31.84	38.27	24.78	38.79	39.66	6.85	7.34	5.64	10.96	9.70	23.77	28.26	22.23	32.76	35.43
Dravidian (4)	kan	0.14	0.79	0.84	1.83	0.79	23.13	33.48	1.65	36.89	39.33	0.02	0.03	0.02	0.35	0.25	14.95	19.35	3.34	37.47	43.46
	mal	0.15	0.35	0.74	3.01	1.38	20.79	34.78	26.20	42.02	46.09	0.04	0.01	0.01	0.97	1.04	11.17	18.23	19.89	36.18	45.78
	tam	14.66	0.77	1.33	3.26	1.88	16.14	29.12	14.19	36.59	40.74	8.91	0.01	0.00	0.70	0.81	9.86	16.16	5.17	33.95	39.09
	tel	17.22	1.66	1.81	2.51	2.02	20.97	35.02	-1.00	40.79	46.50	12.25	0.01	0.07	0.22	0.23	13.40	20.67	-1.00	41.74	48.33
Average		12.18	15.44	11.96	19.79	17.81	31.29	38.03	24.09	38.80	39.83	6.78	6.99	5.37	10.46	9.26	23.23	27.80	21.50	32.98	35.84
Sino-Tibetan (3)	mya	15.07	0.18	0.84	0.80	1.18	3.50	16.01	8.02	30.90	34.06	9.60	0.02	0.06	0.03	0.07	2.57	8.30	7.28	18.66	27.10
	zho_simpl	6.91	15.44	26.14	27.99	25.32	30.52	34.37	26.24	31.07	37.80	15.21	3.46	20.38	20.40	15.08	33.19	33.64	24.98	20.93	39.93
	zho_trad	6.06	12.36	22.78	26.26	24.14	30.05	32.83	-1.00	30.67	35.18	5.63	4.22	11.78	16.30	12.02	24.01	26.49	-1.00	10.97	30.16
Average		12.08	15.23	12.12	19.74	17.78	30.95	37.67	23.64	38.53	39.68	6.89	6.84	5.56	10.52	9.25	23.11	27.63	21.12	32.43	35.73
Other (14)	est	28.08	24.01	6.78	14.74	12.30	40.66	42.21	35.47	36.78	44.49	20.18	8.33	2.71	5.45	4.99	33.71	38.24	35.68	32.73	41.82
	fin	25.78	29.83	8.01	32.24	29.70	35.90	40.17	33.75	35.45	38.99	23.45	11.54	2.86	18.57	14.70	33.38	35.33	33.27	29.97	37.32
	hun	2.32	22.52	8.17	32.46	28.57	36.44	38.58	35.36	35.78	40.86	0.77	6.97	4.34	16.98	13.27	27.37	32.10	35.89	32.27	39.18
	kat	0.32	0.84	1.28	7.15	3.48	12.65	23.78	14.25	29.94	33.96	0.04	0.03	0.01	2.22	3.64	11.13	16.82	3.20	30.67	36.06
	hau	2.91	8.02	6.18	6.33	7.61	16.85	32.20	20.06	39.62	40.67	0.38	1.23	2.05	2.06	3.25	7.87	15.44	13.19	31.79	34.20
	heb	0.40	1.99	1.13	16.29	9.36	38.51	43.97	37.19	41.95	48.88	0.11	0.16	0.09	4.62	4.62	29.04	34.82	37.14	37.57	44.60
	jpn	6.22	19.38	14.18	25.65	23.45	30.57	32.65	26.85	31.67	36.68	17.09	13.84	5.38	21.79	18.95	34.61	35.23	33.27	23.98	42.90
	khm	1.36	0.91	2.71	5.21	5.26	16.03	31.15	21.48	38.68	35.33	0.20	0.04	0.03	0.01	0.12	4.06	7.70	14.44	15.81	25.56
	vie	28.19	18.20	10.63	37.33	32.96	38.93	44.83	38.15	42.16	46.09	27.56	9.45	4.63	26.38	21.71	41.11	41.34	43.24	42.37	48.20
	kor	17.65	4.11	2.59	22.84	22.03	28.56	33.93	27.05	30.55	35.85	9.61	0.19	0.30	11.39	9.06	24.41	26.73	24.42	28.08	31.96
	lao	1.30	2.07	3.53	3.75	3.63	8.81	21.84	19.75	37.49	43.33	0.05	0.08	0.04	0.00	0.00	3.86	11.07	16.83	32.10	30.20
	tha	15.30	1.31	2.93	9.24	8.15	27.49	33.17	27.86	33.84	34.81	16.90	0.03	0.02	1.40	1.83	21.88	25.26	25.47	22.25	38.00
	luo	4.18	7.18	5.84	7.46	8.58	13.08	15.36	-1.00	27.48	-1.00	1.48	1.45	1.56	2.59	2.75	5.61	6.78	-1.00	18.64	-1.00
	mon	1.98	1.05	1.20	3.36	4.41	13.73	22.87	21.13	29.47	38.39	0.11	0.11	0.20	1.25	1.12	5.55	9.69	11.07	21.34	31.72
Average		11.75	14.52	11.18	19.22	17.29	30.21	36.97	23.90	38.05	39.30	7.11	6.42	5.03	10.20	8.96	22.72	27.13	21.42	31.89	35.53

Table 7: Detailed results (BLEU) of our evaluated models on 102 languages.

Language Family	Language	X-to-Eng (COMET)										Eng-to-X (COMET)									
		XGLM-7.5B	OPT-175B	Falcon-7B	LLaMA2-7B	LLaMA2-7B-Chat	ChatGPT	GPT4	M2M-12B	NLLB-1.3B	Google	XGLM-7.5B	OPT-175B	Falcon-7B	LLaMA2-7B	LLaMA2-7B-Chat	ChatGPT	GPT4	M2M-12B	NLLB-1.3B	Google
Indo-European-Germanic (7)	afr	62.96	86.21	80.54	84.93	83.45	89.87	90.33	87.24	88.36	89.73	44.67	69.54	61.85	72.62	68.30	87.00	87.48	85.64	86.56	86.94
	dan	72.74	88.46	84.15	89.10	89.04	90.54	90.67	89.74	90.03	90.76	45.95	80.09	62.31	84.14	79.85	90.79	90.45	88.89	88.92	91.02
	nld	73.32	86.41	83.69	87.42	87.27	88.62	88.64	87.42	88.04	88.39	47.82	81.52	72.40	85.26	82.11	88.87	89.03	87.43	88.39	88.97
	deu	86.13	87.75	86.71	88.08	88.30	89.39	89.61	88.48	88.98	89.50	80.23	78.06	76.54	82.88	79.99	87.95	88.51	85.61	86.26	89.13
	isl	50.24	62.09	54.35	66.68	64.91	85.35	87.13	83.43	85.25	87.50	29.53	32.46	32.01	37.53	42.71	79.22	83.28	81.51	82.92	87.16
	nob	69.85	86.82	81.47	87.55	87.55	89.25	89.47	88.08	86.94	89.07	48.40	81.45	64.72	83.68	80.21	89.86	89.86	87.58	88.37	89.67
	swe	75.42	88.26	0.87	89.32	89.22	90.32	90.50	89.78	89.67	90.54	54.84	80.67	0.69	86.26	82.20	91.09	90.90	88.61	89.74	90.46
Average		70.09	83.71	67.40	84.73	84.25	89.05	89.48	87.74	88.18	89.36	50.21	71.97	52.93	76.05	73.63	87.83	88.50	86.47	87.31	89.05
Indo-European-Romance (4)	cat	86.13	86.73	81.35	88.21	87.86	89.53	89.77	87.86	88.91	89.73	83.64	72.95	61.49	83.44	81.46	88.46	88.33	86.96	87.71	88.31
	fra	86.60	88.73	88.09	88.89	88.67	89.68	89.92	88.61	89.16	89.69	81.81	82.83	84.83	83.97	83.31	88.61	88.39	86.71	87.66	88.39
	glg	82.41	87.06	83.89	86.19	85.90	89.02	88.92	87.42	88.80	88.70	71.62	76.01	71.58	76.02	75.54	87.99	87.92	86.82	87.31	87.52
	ron	63.53	88.49	84.28	89.10	88.75	90.22	90.35	89.34	89.08	90.52	39.24	84.51	63.52	84.76	78.37	90.94	91.11	89.75	89.79	90.83
Average		73.58	85.18	73.58	85.95	85.54	89.25	89.57	87.95	88.47	89.47	57.07	74.55	59.27	78.23	75.82	88.25	88.66	86.87	87.60	88.95
Indo-European-Slavic (12)	bel	48.56	60.07	50.69	70.78	66.92	83.19	84.09	75.99	83.73	84.42	31.23	33.97	31.81	42.44	44.57	79.07	82.53	71.62	85.61	87.53
	bos	57.24	85.93	74.98	87.71	87.66	89.48	89.69	87.83	88.76	89.77	30.09	77.04	45.68	83.56	75.99	89.32	90.69	88.62	90.09	90.53
	bul	85.87	67.92	65.12	87.27	86.11	88.34	88.95	87.89	87.61	88.78	84.67	42.27	38.19	82.37	74.36	89.00	89.98	89.53	89.85	91.32
	hrv	55.20	85.84	74.71	86.48	85.38	88.03	88.27	86.93	86.76	88.55	28.41	78.03	43.08	83.55	73.81	89.41	90.54	87.98	89.08	90.70
	ces	59.62	84.95	79.26	87.53	86.66	88.97	89.17	87.79	88.03	89.63	29.94	65.99	50.67	82.80	75.16	90.57	91.27	88.71	90.18	90.57
	mkd	74.16	64.07	59.09	83.56	81.27	88.06	88.73	87.17	87.59	89.08	59.72	35.44	35.24	68.26	57.43	86.85	87.12	88.96	88.86	89.64
	pol	56.44	84.42	80.90	86.10	85.97	87.26	87.55	86.13	86.63	87.42	31.42	74.55	62.60	83.39	77.50	89.16	89.85	86.65	88.56	90.13
	rus	84.02	73.03	72.87	86.11	86.06	87.36	87.49	86.12	86.88	87.53	81.30	58.53	46.24	83.73	78.29	89.64	89.76	87.65	88.20	89.62
	srp	65.14	59.20	57.21	86.46	86.03	87.91	88.51	86.59	87.24	88.15	43.18	32.69	32.46	82.08	73.98	86.57	89.03	84.72	87.75	89.38
	slk	56.65	82.74	71.95	84.68	83.62	88.40	88.65	86.94	87.56	88.42	30.19	54.34	42.07	62.71	56.39	89.26	89.75	88.06	89.71	90.85
	slv	56.46	80.35	0.70	85.41	85.30	88.11	88.45	86.25	87.07	88.53	28.92	52.41	0.39	77.55	73.24	87.58	89.48	86.88	88.39	90.02
	ukr	71.47	69.11	0.61	86.68	85.81	87.75	88.22	86.00	87.13	87.84	40.25	48.07	0.37	83.14	76.28	88.41	89.78	87.71	88.57	90.26
Average		68.70	79.77	65.11	85.40	84.68	88.46	88.83	86.91	87.75	88.79	49.87	64.06	46.99	77.22	72.66	88.07	88.92	86.64	88.19	89.52
Indo-European-Indo-Aryan (10)	asm	64.09	48.51	48.71	57.33	55.02	79.55	84.57	-	86.19	86.94	30.33	33.47	31.49	34.05	28.06	66.34	72.13	-	82.67	82.90
	ben	83.47	48.51	48.99	66.59	60.13	86.55	88.38	86.22	88.88	89.71	73.13	32.46	29.99	36.90	31.03	75.85	81.77	84.03	86.36	86.30
	guj	47.25	49.48	51.09	52.61	53.72	85.19	89.27	38.57	89.95	90.83	23.10	36.95	34.06	38.91	38.29	73.62	78.90	62.98	87.86	88.33
	hin	85.88	51.18	50.17	80.11	77.50	89.20	90.64	88.76	90.37	90.79	70.40	25.92	26.39	44.57	41.04	76.72	78.61	79.05	81.60	82.40
	mar	62.11	49.64	48.68	65.73	62.18	83.99	87.06	82.09	88.24	88.84	34.50	26.60	22.67	34.53	33.37	59.24	66.22	67.98	74.35	75.59
	npi	73.64	53.66	55.26	74.31	70.13	87.31	90.88	75.24	91.22	91.84	39.40	30.51	24.37	37.98	36.34	70.87	77.36	53.47	81.33	83.59
	ory	47.95	45.52	50.42	52.09	52.18	81.20	87.24	44.71	88.79	89.09	19.94	35.21	32.98	32.16	34.95	60.85	70.37	40.70	83.72	80.50
	pan	47.09	50.42	49.91	52.22	53.45	86.27	89.35	78.38	89.35	89.84	20.70	31.70	29.26	33.02	33.58	70.62	77.34	59.40	84.32	84.69
	snd	46.69	50.29	48.52	57.19	55.35	76.01	81.96	51.82	87.15	87.91	23.50	35.43	26.87	29.17	33.65	53.11	56.16	66.01	80.44	80.30
	urd	81.11	46.53	0.48	68.14	63.22	86.08	88.54	81.15	87.88	88.53	74.95	30.25	0.28	37.90	37.12	77.07	78.66	74.31	82.83	83.20
Average		67.26	70.56	59.08	78.50	77.29	87.15	88.51	82.06	88.07	88.99	47.18	54.30	40.58	64.71	61.17	82.12	84.32	80.64	86.48	87.48
Indo-European-Other (9)	hye	39.23	45.76	45.65	55.59	54.10	76.40	85.17	75.72	88.41	89.26	24.05	35.48	32.49	33.90	34.59	52.00	69.40	66.22	89.80	89.72
	ell	84.35	65.85	60.60	79.99	76.96	87.76	88.33	86.92	87.48	88.26	85.01	46.81	36.63	46.75	43.10	88.13	88.73	88.48	88.31	89.00
	gle	45.61	56.53	55.41	67.80	64.88	84.67	87.30	37.70	84.79	87.84	33.91	34.20	34.79	39.83	42.27	74.21	77.68	33.84	80.53	81.99
	cym	47.05	59.34	0.57	67.67	63.40	87.82	89.58	70.57	87.30	90.02	30.67	32.98	0.36	39.23	38.69	84.59	86.46	70.27	85.56	88.78
	ita	85.44	86.66	87.07	87.65	87.48	88.52	89.02	87.27	88.31	88.82	82.89	82.61	84.83	84.62	82.76	88.56	88.91	87.24	88.05	88.67
	lav	51.23	61.86	56.54	68.26	65.59	87.24	88.14	86.98	86.45	88.50	28.88	34.48	31.68	37.82	38.59	87.26	87.64	87.21	85.79	90.77
	lit	50.67	61.28	59.07	66.68	66.03	87.26	87.54	86.52	85.84	87.43	26.84	33.25	34.31	40.70	40.16	88.09	89.82	87.86	88.62	90.59
	pus	38.23	49.79	49.46	57.78	58.45	73.69	77.52	79.36	85.55	85.99	22.74	32.01	28.86	30.35	33.93	68.69	53.48	68.49	79.48	80.31
	fas	55.82	49.28	49.94	77.43	71.82	87.28	88.21	86.21	87.16	88.50	30.37	28.53	27.85	43.62	39.45	84.42	86.33	84.45	86.24	87.13
Average		64.69	68.21	57.48	76.65	75.22	86.59	88.14	81.05	87.80	88.84	45.77	51.24	39.31	60.29	57.43	81.09	83.60	79.38	86.34	87.47
Austronesian (3)	ind	86.76	84.82	82.85	87.95	87.72	89.74	90.14	88.37	89.21	89.62	85.86	77.45	71.69	85.54	82.96	91.42	91.58	89.34	90.47	91.93
	jav	67.78	66.26	61.53	66.81	67.84	82.77	85.53	77.56	85.71	87.10	52.22	45.55	43.42	57.62	63.86	78.23	81.82	83.30	86.65	86.24
	msa	81.85	83.00	81.69	85.87	85.61	89.36	90.27	88.36	88.62	89.96	81.34	71.91	67.57	81.25	77.82	89.43	89.64	87.84	89.34	89.84
Average		65.63	68.86	58.67	76.89	75.57	86.63	88.17	81.31	87.80	88.84	47.60	52.16	40.75	61.26	58.59	81.44	83.87	79.89	86.51	87.59
Atlantic-Congo (2)	swh	81.12	61.01	0.58	61.34	60.51	87.71	88.25	83.77	86.24	88.16	77.72	33.58	0.32	34.28	39.07	85.51	86.04	83.43	85.65	85.74
	xho	42.89	54.17	0.53	54.10	55.21	71.55	78.59	69.09	81.76	82.72	32.13	34.72	0.37	34.05	37.20	65.00	69.53	68.15	74.26	76.04
Average		65.48	68.38	56.20	76.07	74.81	86.34	87.97	81.09	87.64	88.70	47.91	51.39	39.03	60.10	57.72	81.18	83.61	79.71	86.23	87.31
Afro-Asiatic (4)	amh	44.46	49.38	49.59	49.66	53.22	60.66	81.86	70.20	86.24	88.17	27.68	45.18	34.54	35.00	40.46	52.12	71.50	67.68	85.83	86.34
	ara	81.55	52.45	54.55	78.26	73.84	87.74	88.10	85.66	87.38	88.06	69.71	35.39	35.53	55.73	47.25	86.80	87.05	84.11	86.73	87.92
	orm	44.83	51.23	49.82	49.41	51.21	60.41	65.69	-	77.05	78.82	32.88	42.15	40.13	41.67	44.97	65.94	69.20	-	77.78	80.42
	som	47.21	58.55	0.53	53.54	54.80	72.77	79.40	45.08	81.27	82.84	35.67	44.71	0.37	39.77	40.91	65.52	74.50	54.80	81.03	80.43
Average		64.62	67.17	54.82	74.63	73.52	85.09	87.25	80.23	87.27	88.37	47.40	50.65	38.13	58.77	56.60	80.11	82.98	79.05	85.96	87.03
Turkic (5)	azj	60.35	67.69	57.84	67.91	66.10	86.59	87.48	61.04	87.30	88.04	38.01	32.40	32.36	40.51	43.85	81.78	83.37	78.26	87.09	87.71
	kaz	56.44	51.68	53.62	61.98	63.24	81.67	86.05	42.15	87.40	88.91	27.24	41.62	33.37	32.24	35.51	66.34	71.74	64.63	88.95	90.43
	kir	53.59	50.36	52.88	59.64	61.46	78.36	82.67	-	85.99	86.77	24.41	40.42	33.53	34.66	36.22	58.30	63.59	-	87.96	88.10
	tur	84.23	83.41	0.66	80.46	79.76	89.85	90.33	87.89	88.90	89.91	74.73	69.49	0.37	54.42	53.06	88.64	89.53	86.10	88.81	90.40
	uzb	53.87	59.24	0.57	59.77	60.19	83.75	87.96	43.01	87.97	89.03	37.21	40.09	0.35	36.44	41.21	78.79	84.27	43.84	89.52	90.27
Average		64.36	66.75	52.88	73.86	72.86	84.99	87.22	78.59	87.30	88.38	46.77	50.12	36.51	57.06	55.29	79.64	82.58	78.23	86.19	87.24
Dravidian (4)	kan	44.69	43.02	48.00	50.82	50.97	82.92	87.07	39.28	88.09	88.20	19.74	33.61	29.90	33.98	34.83	69.49	76.68	54.48	84.95	85.55
	mal	44.84	44.51	48.20	54.08	53.89	83.74	88.05	83.55	89.91	90.74	25.76	31.41	30.71	33.47	36.21	60.57	73.01	77.38	86.68	88.98
	tam	79.12	41.05	0.47	56.12	54.10	79.64	85.59	68.48	87.24	88.07	76.92	32.42	0.32	34.02	36.85	63.35	76.50	54.90	88.45	89.03
	tel	79.16	47.46	0.50	51.66	52.87	81.71	86.93	-	88.44	89.39	70.17	32.60	0.31	34.64	34.70	65.56	74.40	-	85.19	87.43
Average		64.20	65.23	50.97	72.48	71.53	84.79	87.20	77.80	87.37	88.43	46.86	48.95	35.10	55.52	53.98	78.64	82.08	77.37	86.19	87.28
Sino-Tibetan (3)	mya	79.89	41.28	48.84	51.91	52.20	61.03	77.98	57.30	86.86	87.42	80.82	41.01	36.02	29.51	37.45	50.43	65.73	64.16	84.68	87.38
	zho_simpl	49.06	78.34	84.94	85.94	85.69	87.31	87.88	85.40	86.16	88.11	74.82	59.17	83.28	83.94	78.04	88.88	88.73	83.49	78.56	89.05
	zho_trad	46.85	76.33	83.26	85.50	84.73	87.21	87.58	-	86.47	87.46	66.84	64.03	80.92	84.54	79.73	88.82	88.89	-	78.98	89.12
Average		63.93	65.24	51.99	72.57	71.66	84.49	87.07	77.58	87.33	88.39	48.16	49.23	36.61	56.02	54.51	78.52	82.04	77.25	85.93	87.33
Other (13)	est	86.04	81.09	58.12	70.45	68.72	89.64	89.72	87.81	87.98	89.94	82.72	57.08	33.17	41.19	43.79	90.88	92.05	88.69	89.80	91.21
	fin	86.66	86.74	63.92	88.33	87.23	90.24	90.33	89.11	89.16	90.04	86.71	71.23	37.15	83.95	75.89	92.05	92.42	89.46	89.95	91.45
	hun	42.14	80.15	60.25	86.13	85.86	88.52	88.96	87.14	87.24	88.79	25.15	55.31	33.98	79.78	74.66	87.77	89.50	87.72	88.59	89.95
	kat	43.25	41.69	46.15	63.34	60.86	76.21	83.55	71.85	86.46	87.93	25.02	32.74	27.44	34.01	37.20	50.80	65.45	40.66	83.36	87.34
	hau	48.82	56.00	54.60	53.67	55.26	69.81	79.33	66.06	83.02	83.18	36.01	38.35	35.84	34.99	38.42	58.87	70.78	63.83	80.89	81.31
	heb	40.58	49.75	46.81	70.79	65.39	87.35	88.85	86.68	87.65	89.22	22.30	31.75	27.66	41.54	40.40	82.26	84.68	86.70	87.04	88.61
	jpn	53.32	82.82	79.06	86.41	86.47	88.33	88.86	87.19	88.04	88.55	81.81	78.90	61.42	84.84	82.02	91.24	90.88	88.02	88.70	92.37
	khm	45.09	39.56	50.21	56.63	57.10	77.48	85.48	72.07	87.13	85.77	23.92	37.39	34.14	28.23	32.00	48.38	58.96	71.75	79.20	82.19
	vie	83.91	71.02	0.67	86.61	85.51	88.01	89.02	86.74	87.33	88.49	84.30	58.57	0.41	80.27	74.81	88.37	89.14	87.85	88.06	89.79
	kor	82.74	57.11	54.95	85.39	84.83	88.17	88.69	86.35	87.13	88.66	72.09	35.99	30.50	75.21	68.76	88.48	89.21	86.21	87.94	89.01
	lao	48.13	46.72	51.74	53.14	54.60	67.62	78.00	73.99	86.31	88.28	28.63	38.31	32.85	31.09	31.31	43.60	54.32	64.97	83.50	81.58
	tha	75.17	49.20	0.55	69.76	66.38	86.70	88.47	85.58	86.97	86.85	78.17	31.91	0.29	42.34	42.78	82.92	84.65	82.47	83.03	88.12
	mon	49.71	48.31	48.88	54.32	56.34	74.20	81.49	79.36	84.82	87.39	23.58	41.57	31.63	34.10	35.55	60.45	72.59	73.11	85.84	88.76
Average		63.33	64.48	51.20	72.33	71.43	84.15	86.92	78.30	87.25	88.31	48.75	48.82	35.43	55.54	54.10	77.80	81.62	77.35	85.92	87.42

Table 8: Detailed results (COMET) of our evaluated models.

Language Family	Language	X⇒Eng (SEScore)									
		XGLM-7.5B	OPT-175B	Falcon-7B	LLaMA2-7B	LLaMA2-7B-Chat	ChatGPT	GPT4	M2M-12B	NLLB-1.3B	Google
Indo-European-Germanic (8)	afr	-13.42	-3.14	-6.72	-3.59	-4.30	-0.48	-0.19	-1.77	-1.67	-0.24
	dan	-10.98	-3.18	-5.76	-2.66	-2.68	-1.35	-1.15	-1.98	-1.65	-0.93
	nld	-10.66	-4.76	-6.31	-4.25	-4.53	-3.66	-3.59	-4.17	-3.59	-3.41
	deu	-4.44	-3.41	-4.56	-3.17	-3.23	-2.21	-2.04	-2.74	-2.37	-1.93
	isl	-20.12	-15.12	-18.05	-13.19	-14.49	-4.99	-4.09	-5.59	-5.04	-3.16
	ltz	-12.83	-11.46	-13.68	-10.39	-11.89	-3.14	-2.12	-4.06	-2.01	-1.52
	nob	-12.37	-3.96	-7.10	-3.60	-3.54	-2.52	-2.33	-2.85	-3.58	-2.24
	swe	-9.42	-2.96	-4.52	-2.40	-2.57	-1.78	-1.80	-2.02	-2.33	-1.34
Average		-11.78	-6.00	-8.34	-5.41	-5.90	-2.52	-2.16	-3.15	-2.78	-1.85
Indo-European-Romance (8)	ast	-6.71	-5.74	-6.92	-5.61	-5.95	-2.82	-2.42	-4.02	-3.47	-
	cat	-4.07	-3.69	-7.42	-2.93	-3.18	-2.00	-1.77	-2.66	-2.14	-1.34
	fra	-4.30	-2.97	-3.39	-2.87	-3.04	-2.08	-1.82	-2.62	-2.61	-1.84
	glg	-6.40	-4.17	-6.25	-4.36	-4.80	-2.64	-2.68	-3.32	-2.48	-2.67
	oci	-5.89	-4.70	-6.00	-4.11	-5.23	-1.48	-0.51	-2.58	-0.88	-
	por	-3.53	-2.66	-3.39	-2.35	-2.76	-1.17	-1.39	-2.18	-1.82	-1.40
	ron	-15.54	-3.15	-5.78	-2.76	-3.06	-2.10	-1.92	-2.35	-2.31	-1.30
	spa	-5.88	-5.02	-5.41	-4.78	-5.13	-4.14	-4.12	-4.92	-4.57	-4.18
Average		-9.16	-5.01	-6.95	-4.56	-5.02	-2.41	-2.12	-3.11	-2.66	-1.96
Indo-European-Slavic (12)	bel	-20.43	-17.35	-19.92	-12.05	-13.99	-6.79	-6.12	-9.50	-6.25	-5.86
	bos	-17.40	-4.83	-10.39	-3.74	-4.31	-2.46	-2.28	-3.29	-2.98	-1.86
	bul	-4.33	-13.68	-15.29	-3.58	-4.49	-2.80	-2.25	-2.78	-2.93	-1.67
	hrv	-18.09	-5.29	-10.96	-4.60	-4.81	-3.73	-3.60	-3.96	-4.18	-3.24
	ces	-17.19	-5.39	-9.01	-3.88	-4.55	-2.80	-2.77	-3.54	-3.42	-2.18
	mkd	-9.69	-16.10	-17.54	-5.61	-6.86	-3.02	-2.34	-3.28	-3.04	-1.83
	pol	-17.76	-5.89	-7.88	-4.94	-5.09	-4.32	-3.91	-4.62	-4.24	-3.59
	rus	-5.97	-11.16	-11.51	-4.64	-4.76	-3.83	-3.61	-4.47	-3.68	-3.17
	srp	-13.95	-17.28	-18.10	-3.88	-4.43	-3.05	-2.56	-3.70	-3.24	-2.36
	slk	-17.84	-6.57	-11.48	-5.78	-6.38	-3.32	-2.93	-3.67	-3.54	-2.75
	slv	-17.77	-7.74	-12.92	-5.21	-5.41	-3.45	-3.31	-4.17	-3.90	-3.10
	ukr	-11.07	-12.44	-16.56	-3.39	-3.94	-3.01	-2.40	-3.57	-2.97	-1.98
Average		-11.36	-7.28	-9.74	-4.80	-5.34	-2.90	-2.57	-3.59	-3.10	-2.35
Indo-European-Indo-Aryan (10)	asm	-17.62	-22.44	-21.75	-19.23	-21.61	-10.07	-7.23	-	-5.46	-5.02
	ben	-8.64	-22.29	-21.85	-16.07	-19.59	-6.90	-4.90	-5.50	-4.06	-3.18
	guj	-22.60	-22.48	-21.66	-21.04	-22.00	-7.84	-4.68	-23.21	-3.36	-2.52
	hin	-6.78	-21.75	-21.65	-9.46	-11.55	-4.05	-2.65	-3.56	-2.54	-1.69
	mar	-17.74	-22.28	-21.98	-16.22	-19.49	-7.44	-4.84	-6.94	-3.60	-2.75
	npi	-15.26	-21.15	-20.97	-14.08	-17.05	-6.54	-2.94	-10.52	-2.58	-1.41
	ory	-22.89	-22.85	-21.75	-21.22	-22.60	-10.30	-5.71	-22.74	-3.91	-3.52
	pan	-22.96	-22.04	-21.63	-20.72	-22.01	-6.20	-3.30	-8.54	-3.01	-2.27
	snd	-21.45	-21.71	-21.57	-18.93	-21.03	-11.57	-7.19	-17.98	-3.25	-2.66
	urd	-8.52	-22.49	-21.67	-14.55	-17.63	-5.52	-3.43	-6.98	-3.56	-2.99
Average		-12.70	-11.19	-12.88	-8.05	-9.05	-4.15	-3.13	-5.58	-3.22	-2.47
Indo-European-Other (11)	hye	-23.28	-22.38	-22.07	-18.87	-21.02	-11.09	-5.55	-8.90	-3.47	-2.53
	ell	-5.76	-14.88	-16.79	-7.61	-9.70	-3.66	-3.29	-4.15	-3.57	-2.77
	gle	-20.94	-17.48	-17.96	-12.56	-14.43	-4.07	-2.30	-21.85	-3.03	-1.36
	cym	-20.63	-17.06	-17.53	-12.22	-15.38	-1.95	-0.45	-8.77	-1.78	0.24
	ita	-5.34	-4.82	-4.90	-4.12	-3.92	-3.60	-3.34	-4.02	-3.53	-3.07
	lav	-19.87	-16.43	-17.88	-13.19	-15.73	-4.44	-3.61	-4.12	-4.38	-2.79
	lit	-20.00	-16.62	-17.12	-13.55	-14.81	-4.49	-3.96	-4.27	-4.67	-3.27
	pus	-23.23	-21.32	-21.25	-18.69	-20.25	-12.83	-10.00	-7.44	-4.25	-4.08
	fas	-19.32	-21.16	-20.69	-10.03	-12.97	-4.29	-3.54	-4.59	-4.17	-2.77
	ckb	-22.58	-22.60	-21.94	-20.09	-21.42	-13.33	-8.76	-	-	-22.36
	tgk	-21.03	-21.16	-20.90	-18.74	-20.04	-10.41	-6.04	-	-4.64	-3.64
Average		-13.97	-12.68	-14.05	-9.30	-10.48	-4.73	-3.46	-5.97	-3.33	-2.93
Austronesian (6)	ceb	-18.67	-9.30	-13.22	-11.12	-12.24	-4.20	-2.08	-8.10	-2.67	-1.09
	tgl	-18.04	-5.94	-10.11	-7.18	-8.10	-2.25	-1.53	-5.82	-2.43	-1.53
	ind	-4.84	-6.02	-8.15	-4.01	-4.23	-2.56	-2.33	-3.13	-2.90	-2.28
	jav	-14.95	-14.93	-16.14	-14.04	-14.98	-6.25	-3.94	-6.97	-3.56	-2.83
	msa	-6.84	-6.73	-7.99	-4.95	-5.24	-2.75	-1.68	-2.82	-2.74	-1.53
	mri	-21.00	-17.56	-18.19	-16.36	-18.08	-8.88	-6.64	-	-6.51	-6.09
Average		-13.98	-12.39	-13.86	-9.33	-10.48	-4.70	-3.42	-5.91	-3.34	-2.88
Atlantic-Congo (14)	lug	-20.90	-18.15	-18.81	-18.11	-18.81	-14.28	-10.65	-19.72	-8.06	-7.53
	ibo	-21.95	-19.14	-19.17	-18.70	-19.60	-15.44	-11.98	-12.60	-6.29	-6.56
	kea	-14.56	-9.88	-13.84	-10.94	-11.97	-3.44	-1.64	-	-2.92	-
	kam	-19.35	-17.92	-19.02	-18.11	-18.57	-15.87	-14.95	-	-10.85	-
	lin	-19.99	-17.51	-18.27	-17.44	-18.70	-14.65	-11.86	-17.78	-6.37	-6.55
	nso	-20.19	-17.77	-18.22	-17.61	-19.05	-13.09	-7.08	-17.27	-4.41	-
	nya	-20.06	-17.96	-18.32	-18.05	-18.51	-11.50	-8.07	-	-6.96	-6.49
	sna	-21.21	-18.09	-18.80	-18.03	-19.02	-12.83	-8.95	-	-7.00	-7.08
	swh	-6.38	-16.24	-17.33	-15.75	-17.27	-2.42	-1.59	-4.08	-2.85	-1.36
	umb	-21.73	-19.15	-19.69	-19.03	-20.22	-17.99	-17.05	-	-12.75	-
	wol	-20.21	-17.34	-18.81	-17.96	-18.85	-16.39	-13.95	-18.38	-10.07	-
	xho	-22.28	-18.37	-19.12	-18.18	-18.82	-10.39	-6.29	-8.70	-4.49	-3.58
	yor	-20.94	-19.45	-19.33	-19.56	-20.00	-14.89	-11.11	-19.97	-8.74	-8.86
	zul	-22.18	-19.63	-19.41	-18.74	-19.36	-10.17	-5.55	-8.98	-4.50	-3.72
Average		-15.08	-13.45	-14.79	-11.01	-12.11	-6.26	-4.62	-7.15	-4.07	-3.30
Afro-Asiatic (6)	amh	-22.90	-22.15	-21.98	-21.75	-21.78	-19.81	-8.12	-12.19	-5.34	-3.84
	ara	-7.46	-20.31	-18.95	-8.72	-11.46	-3.72	-3.02	-4.33	-3.21	-2.36
	ful	-19.87	-18.57	-18.88	-18.57	-19.21	-17.33	-16.43	-21.80	-	-
	mlt	-19.57	-14.71	-15.96	-11.51	-12.84	-2.64	-0.69	-	-0.36	0.17
	orm	-22.10	-20.04	-19.91	-19.80	-20.86	-17.83	-14.09	-	-7.58	-6.36
	som	-21.21	-17.68	-19.32	-19.31	-19.78	-11.63	-7.32	-19.33	-5.79	-5.08
Average		-15.38	-13.89	-15.14	-11.45	-12.55	-6.73	-4.91	-7.60	-4.10	-3.31
Turkic (5)	azj	-18.08	-16.01	-18.91	-15.32	-17.02	-7.13	-6.08	-15.70	-5.94	-5.26
	kaz	-19.54	-20.68	-20.30	-17.39	-17.92	-8.76	-5.99	-20.87	-4.75	-3.48
	kir	-20.36	-21.28	-20.15	-17.95	-18.49	-11.35	-8.38	-	-5.96	-5.36
	tur	-7.45	-8.08	-15.06	-9.06	-10.21	-3.59	-2.74	-4.05	-3.84	-2.72
	uzb	-20.32	-18.90	-18.89	-17.79	-18.70	-7.34	-4.32	-20.52	-3.94	-2.85
Average		-15.50	-14.08	-15.36	-11.71	-12.79	-6.79	-4.95	-8.05	-4.15	-3.36
Dravidian (4)	kan	-22.74	-22.73	-22.14	-21.22	-22.52	-8.51	-5.29	-22.71	-4.08	-3.73
	mal	-22.96	-22.72	-21.88	-19.81	-22.07	-9.12	-4.77	-6.48	-3.15	-2.41
	tam	-10.36	-22.89	-22.09	-18.83	-21.50	-10.17	-6.08	-11.87	-4.31	-3.49
	tel	-10.00	-21.96	-21.55	-20.87	-21.75	-9.24	-5.25	-	-3.49	-2.65
Average		-15.54	-14.49	-15.67	-12.11	-13.23	-6.91	-4.97	-8.29	-4.13	-3.34
Sino-Tibetan (3)	mya	-11.16	-22.86	-22.11	-21.43	-22.57	-21.11	-11.22	-17.98	-5.48	-4.77
	zho_simpl	-23.28	-10.92	-7.05	-6.14	-6.41	-5.14	-4.53	-5.88	-5.58	-3.80
	zho_trad	-23.78	-11.81	-7.94	-6.41	-7.04	-5.03	-4.62	-	-5.44	-4.34
Average		-15.68	-14.51	-15.56	-12.08	-13.19	-7.03	-5.03	-8.39	-4.18	-3.38
Other (14)	est	-5.69	-8.04	-17.97	-13.04	-14.73	-3.16	-3.14	-3.99	-4.34	-2.54
	fin	-6.27	-5.99	-16.63	-4.99	-5.39	-3.81	-3.30	-4.53	-4.55	-3.41
	hun	-21.72	-8.73	-17.42	-5.54	-5.94	-4.10	-3.74	-4.60	-4.86	-3.72
	kat	-22.75	-22.83	-22.02	-17.02	-19.63	-11.63	-7.29	-11.45	-5.48	-4.62
	hau	-20.67	-18.03	-18.60	-18.46	-18.69	-12.65	-6.87	-11.24	-4.75	-4.49
	heb	-22.76	-20.91	-21.15	-11.53	-15.08	-3.69	-2.51	-3.52	-3.12	-1.98
	jpn	-21.76	-8.96	-11.29	-6.90	-6.86	-4.93	-4.55	-5.42	-4.89	-4.20
	khm	-22.43	-22.92	-21.48	-19.24	-20.46	-12.80	-6.40	-10.90	-4.58	-4.83
	vie	-6.51	-12.47	-14.21	-4.70	-5.90	-3.77	-2.90	-4.06	-3.72	-2.62
	kor	-8.77	-19.08	-19.67	-7.17	-7.83	-5.60	-4.82	-5.87	-5.24	-4.54
	lao	-21.93	-21.41	-20.84	-20.81	-21.52	-17.31	-10.60	-10.21	-4.79	-3.66
	tha	-11.91	-21.89	-19.89	-14.48	-17.15	-5.99	-4.44	-5.98	-4.86	-4.85
	luo	-20.23	-18.64	-18.92	-18.85	-19.16	-17.27	-16.27	-	-7.91	-
	mon	-20.93	-21.89	-21.66	-20.03	-20.07	-12.41	-8.19	-8.04	-6.03	-3.92
Average		-15.82	-14.80	-15.99	-12.22	-13.32	-7.23	-5.17	-8.17	-4.28	-3.44

Table 9: Detailed results (SEScore) of our evaluated models.

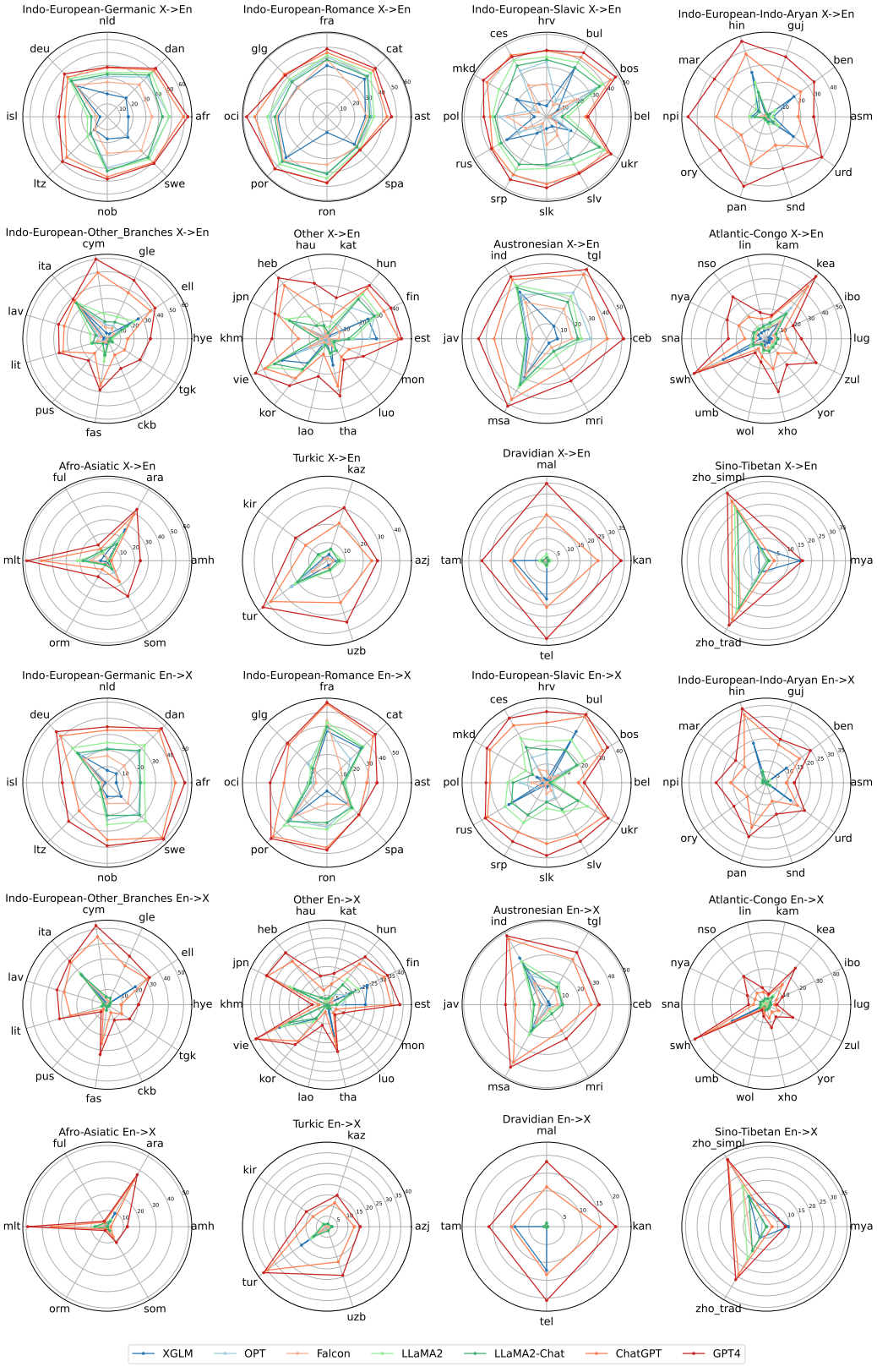

Figure 8: Comparison results between our evalutated LLMs on different language families.

Language	ISO 639-1	ISO 639-2/T	Language family	Language	ISO 639-1	ISO 639-2/T	Language family
Afrikaans	af	afr	Indo-European-Germanic	Latvian	lv	lav	Indo-European-Other
Amharic	am	amh	Afro-Asiatic	Lingala	ln	lin	Atlantic-Congo
Arabic	ar	ara	Afro-Asiatic	Lithuanian	lt	lit	Indo-European-Other
Armenian	hy	hye	Indo-European-Other	Luo	luo	luo	Other
Assamese	as	asm	Indo-European-Indo-Aryan	Luxembourgish	lb	ltz	Indo-European-Germanic
Asturian	ast	ast	Indo-European-Romance	Macedonian	mk	mkd	Indo-European-Slavic
Azerbaijani	az	azj	Turkic	Malay	ms	msa	Austronesian
Belarusian	be	bel	Indo-European-Slavic	Malayalam	ml	mal	Dravidian
Bengali	bn	ben	Indo-European-Indo-Aryan	Maltese	mt	mlt	Afro-Asiatic
Bosnian	bs	bos	Indo-European-Slavic	Maori	mi	mri	Austronesian
Bulgarian	bg	bul	Indo-European-Slavic	Marathi	mr	mar	Indo-European-Indo-Aryan
Burmese	my	mya	Sino-Tibetan	Mongolian	mn	mon	Other
Catalan	ca	cat	Indo-European-Romance	Nepali	ne	npi	Indo-European-Indo-Aryan
Cebuano	ceb	ceb	Austronesian	Northern Sotho	ns	nso	Atlantic-Congo
Chinese (Simpl)	zh	zho_simpl	Sino-Tibetan	Norwegian	no	nob	Indo-European-Germanic
Chinese (Trad)	zhtrad	zho_trad	Sino-Tibetan	Nyanja	ny	nya	Atlantic-Congo
Croatian	hr	hrv	Indo-European-Slavic	Occitan	oc	oci	Indo-European-Romance
Czech	cs	ces	Indo-European-Slavic	Oriya	or	ory	Indo-European-Indo-Aryan
Danish	da	dan	Indo-European-Germanic	Oromo	om	orm	Afro-Asiatic
Dutch	nl	nld	Indo-European-Germanic	Pashto	ps	pus	Indo-European-Other
English	en	eng	Indo-European-Germanic	Persian	fa	fas	Indo-European-Other
Estonian	et	est	Other	Polish	pl	pol	Indo-European-Slavic
Tagalog	tl	tgl	Austronesian	Portuguese	pt	por	Indo-European-Romance
Finnish	fi	fin	Other	Punjabi	pa	pan	Indo-European-Indo-Aryan
French	fr	fra	Indo-European-Romance	Romanian	ro	ron	Indo-European-Romance
Fulah	ff	ful	Afro-Asiatic	Russian	ru	rus	Indo-European-Slavic
Galician	gl	glg	Indo-European-Romance	Serbian	sr	srp	Indo-European-Slavic
Luganda	lg	lug	Atlantic-Congo	Shona	sn	sna	Atlantic-Congo
Georgian	ka	kat	Other	Sindhi	sd	snd	Indo-European-Indo-Aryan
German	de	deu	Indo-European-Germanic	Slovak	sk	slk	Indo-European-Slavic
Greek	el	ell	Indo-European-Other	Slovenian	sl	slv	Indo-European-Slavic
Gujarati	gu	guj	Indo-European-Indo-Aryan	Somali	so	som	Afro-Asiatic
Hausa	ha	hau	Other	Kurdish	ku	ckb	Indo-European-Other
Hebrew	he	heb	Other	Spanish	es	spa	Indo-European-Romance
Hindi	hi	hin	Indo-European-Indo-Aryan	Swahili	sw	swh	Atlantic-Congo
Hungarian	hu	hun	Other	Swedish	sv	swe	Indo-European-Germanic
Icelandic	is	isl	Indo-European-Germanic	Tajik	tg	tgk	Indo-European-Other
Igbo	ig	ibo	Atlantic-Congo	Tamil	ta	tam	Dravidian
Indonesian	id	ind	Austronesian	Telugu	te	tel	Dravidian
Irish	ga	gle	Indo-European-Other	Thai	th	tha	Other
Italian	it	ita	Indo-European-Other	Turkish	tr	tur	Turkic
Japanese	ja	jpn	Other	Ukrainian	uk	ukr	Indo-European-Slavic
Javanese	jv	jav	Austronesian	Umbundu	umb	umb	Atlantic-Congo
Kabuverdianu	kea	kea	Atlantic-Congo	Urdu	ur	urd	Indo-European-Indo-Aryan
Kamba	kam	kam	Atlantic-Congo	Uzbek	uz	uzb	Turkic
Kannada	kn	kan	Dravidian	Vietnamese	vi	vie	Other
Kazakh	kk	kaz	Turkic	Welsh	cy	cym	Indo-European-Other
Khmer	km	khm	Other	Wolof	wo	wol	Atlantic-Congo
Korean	ko	kor	Other	Xhosa	xh	xho	Atlantic-Congo
Kyrgyz	ky	kir	Turkic	Yoruba	yo	yor	Atlantic-Congo
Lao	lo	lao	Other	Zulu	zu	zul	Atlantic-Congo

Table 10: For each language, we list its language name, ISO code and language family.