UvA-MT's Participation in the WMT24 General Translation Shared Task

Shaomu Tan **Di Wu** **David Stap** **Seth Aycock** **Christof Monz**
Language Technology Lab
University of Amsterdam
{s.tan, d.wu, d.stap, s.aycock, c.monz}@uva.nl

Abstract

Fine-tuning Large Language Models (FT-LLMs) with parallel data has emerged as a promising paradigm in recent machine translation research. In this paper, we explore the effectiveness of FT-LLMs and compare them to traditional encoder-decoder Neural Machine Translation (NMT) systems under the WMT24 general MT shared task for English to Chinese direction. We implement several techniques, including Quality Estimation (QE) data filtering, supervised fine-tuning, and post-editing that integrate NMT systems with LLMs.

We demonstrate that fine-tuning LLaMA2 on a high-quality but relatively small bitext dataset (100K) yields COMET results comparable to much smaller encoder-decoder NMT systems trained on over 22 million bitexts. However, this approach largely underperforms on surface-level metrics like BLEU and ChrF. We further control the data quality using the COMET-based quality estimation method. Our experiments show that 1) filtering low COMET scores largely improves encoder-decoder systems, but 2) no clear gains are observed for LLMs when further refining the fine-tuning set. Finally, we show that combining NMT systems with LLMs via post-editing generally yields the best performance for the WMT24 official test set.

1 Introduction

Generative Large Language Models (LLMs) have demonstrated significant capabilities across various English-centric NLP tasks (Zhang et al., 2022; Touvron et al., 2023a,b). However, they often underperform in multilingual contexts, particularly with low-resource languages (Hendy et al., 2023; Stap and Araabi, 2023; Wang et al., 2023). To enhance the multilingual proficiency of LLMs, recent studies have explored several strategies, including vocabulary expansion (Lin et al., 2022; Liang et al., 2023; Yang et al., 2023), continual training on multilingual data (Le Scao et al., 2023; Dubey et al., 2024; Xu et al., 2024a), and instruction tuning (Zhu et al., 2023; Alves et al., 2024; Stap et al., 2024). These approaches have collectively improved LLM performance on a variety of multilingual tasks, such as understanding (Lai et al., 2023), reasoning (Ponti et al., 2020; Shi et al., 2022), summarization (Hasan et al., 2021; Bhattacharjee et al., 2023), and machine translation (Kocmi et al., 2023).

Fine-tuning Large Language Models (FT-LLMs) with parallel data largely enhances translation capabilities, but such approach relies heavily on high-quality parallel data. For instance, prior research often uses development and test datasets like WMT and Flores (Alves et al., 2023; Xu et al., 2024a; Li et al., 2024) for the training, limiting the scalability to a broader range of languages. In this paper, we explore the feasibility of mining high-quality bi-texts from open-source corpora like OPUS. We utilize COMET (Rei et al., 2020), an automated Quality Estimation (QE) tool, to score sentences in the WMT-24 Constraint track. Unlike Peter et al. (2023), who found that selecting the highest quality sentences using COMET improves translation quality, our findings show that while this QE-based data filtering does not provide clear benefits for LLMs when refining fine-tuning datasets, it significantly enhances the performance of NMT systems when applied to filter training samples with low COMET scores.

Recent studies show that LLMs fine-tuned with MT data can rival state-of-the-art NMT models like NLLB (Costa-jussà et al., 2022). However, such comparisons may be unfair, as NMT models like NLLB typically support a broader range of languages. For example, ALMA-13b (Xu et al., 2024a) outperforms NLLB-54b (Costa-jussà et al., 2022) despite targeting only eight language pairs versus 200. Additionally, expanding languages in multilingual models often causes interference that degrades performance (Tan et al., 2024; Shaham et al., 2023). In this paper, we focus exclusively on

Proceedings of the Ninth Conference on Machine Translation, pages 176–184
November 15-16, 2024 ©2024 Association for Computational Linguistics

the English-to-Chinese translation direction[1], investigating how FT-LLMs compare to NMT models trained from scratch using the same parallel data source. Specifically, we use the full WMT-24 constraint track data to train an encoder-decoder NMT model, and we fine-tune LLaMA2-7B on a selected high-quality subset of up to 300K sentences. we found that, despite fine-tuned LLama2-7B being 17 times larger, it yields comparable COMET scores and worse scores for BLEU and ChrF.

While small NMT systems are resource-efficient in production, LLMs in practice, generate less literal translations (Vilar et al., 2023). In this paper, we integrate NMT and LLM systems by prompting LLMs to post-edit (PED) NMT outputs. Additionally, we implement a QE-guided PED system that selects the final outputs based on the higher QE score, as determined by COMET, between NMT and post-edited outputs. Our experiments show that the QE-guided PED system delivers the best performance on the WMT24 en-zh official test set, improving ChrF up to +3.7 over pure NMT outputs and +2.1 than direct translations by LLMs. Surprisingly, this approach brings negative performance gains on the Flores-devtest and Ntrex.

2 Data Preprocessing

In this section, we provide an overview of the data sources and the cleaning strategy. We use all the available data from the constrained track of the WMT-24 shared task for all three directions in which we participate, including English→Chinese, English→Japanese, and Japanese→Chinese. Following Wu et al. (2023), we perform a thorough preprocessing phase involving three key steps to enhance the data quality, as outlined below.

- Character-level Cleaning
  - Deescaping special characters in XML.
  - Removing non-printable characters.
  - Segmenting Chinese sentences with Jieba[2] and tokenizing Japanese data using KyTea (Neubig et al., 2011).
- Sentence-level Cleaning
  - Filtering out sentences longer than 256 tokens.
  - Eliminating sentences where over 75% of the words on both the source and target sides are identical.
  - Removing sentences with a source-to-target token ratio exceeding 3.0.
  - Eliminating duplicated sentences.
- Language-level Cleaning
  - Removing off-target sentences using the FastText language identification tool (Joulin et al., 2016).
  - Excluding sentences exhibiting one-to-many or many-to-one mappings, for example, a single source sentence having multiple different target sentences.

In specific, we use the Moses toolkit[3](Koehn et al., 2007) for all procedures in cleaning step 1 and use FastText (Joulin et al., 2016) for the language identification step. As shown in Table 1 (Cleaned), we removed 29%, 22%, and 45% of the data for en→zh, en→ja, and ja→zh directions.

Directions	Raw	Cleaned	QE-filtered
en→zh	55,346,004	39,354,051	22,606,804
en→ja	33,875,162	26,415,631	14,507,351
ja→zh	22,642,553	12,560,471	6,679,265

Table 1: Number of parallel sentences for three datasets.

3 Systems

3.1 NMT Systems

MMT baseline In this section, we describe the backbone architecture and adjustments made to our baseline systems. We train a multilingual-Transformer-large (mT-large) model for all three en→zh, en→ja, ja→zh directions. The mT-large is a 12-layer Transformer (Vaswani et al., 2017) architecture with specific modifications, including prenorm for both the encoder and decoder, and layernorm for embedding. To enhance stability and performance, we tie the parameters of encoder embedding, decoder embedding, and decoder output. We also introduce dropout and attention dropout with a probability of 0.1, along with label smoothing at a rate of 0.1. In addition, to specify the translation directions, we prepend the source language tags in the source, and target language tags in the target side, e.g.: en2zh.

[1] We investigate FT-LLM for en-zh, and explore the data filtering for en-zh, en-ja, and ja-zh diretcions.

[2] https://github.com/fxsjy/jieba

[3] https://github.com/moses-smt/mosesdecoder/

Similar to the approach described by Vaswani et al. (2017), we employ the Adam optimizer with a learning rate of 5e-4, implementing an inverse square root learning rate schedule with 4,000 warmup steps. We set the maximum number of tokens to 10,240, with gradient accumulation every 21 steps to facilitate large-batch training in Tang et al. (2021). We train all of our systems with 4 NVIDIA A6000 Gpus, and to expedite the training process, we conducted all experiments using half-precision training (FP16). Additionally, we save checkpoints every 2000 steps and implement early stopping based on perplexity, with a patience of 5 epochs.

Quality-Estimation Filtering. Due to data scarcity in the machine translation community, a large amount of Machine Translation data is mined from web-crawled data such as CCAligned (El-Kishky et al., 2020). Nonetheless, recent research found that there are many misaligned data exist in such web-crawled datasets, which impair performance when training models on it (Khayrallah and Koehn, 2018; Ranathunga et al., 2024). In addition, incorrect language and non-linguistic contents could affect the model in generating off-target or hallucinated outputs (Kreutzer et al., 2022). Similarly, recent studies on instruction fine-tuning of LLMs have shown that increasing data quality is more effective than data quantity (Du et al., 2023; Pan et al., 2024; Zhou et al., 2024), especially in inducing instruction-related capabilities (Xia et al., 2024). Additionally, Peter et al. (2023) shows that using QE metrics is not as effective at detecting translation noises like untranslated sentences, but is much better at identifying more fine-grained problems in the data, like small translation or grammatical errors.

Motivated by that, we investigate the feasibility of extracting high-quality parallel data using an automated Quality Estimation (QE) tool. We utilize the COMETKiwi model and apply this data-filtering phase to the cleaned data that we discussed in Section 2. Figure 1 presents the COMET score distributions for three directions. We found that for both English→Chinese and English→Japanese, the distributions are quite similar, that is, nearly half of the data falls into the poor quality range (0-80% Comet scores). For Japanese→Chinese, approximately half dataset ranges from 0% to 65% of COMET score. According to this observation, we filtered out parallel data that has smaller than 80% Comet scores for both English→Chinese and English→Japanese, and set the threshold at 65% for Japanese→Chinese. As a result, we show the number of parallel sentences after Quality Estimation filtering in Table 1.

Directional Fine-tuning. Lastly, to encourage the MMT model to gradually narrow down the data distribution to focus on task-specific data, we further fine-tune the MMT model on direction-specific data. Note that the direction-specific data, i.e., En → Zh, En → Ja, and Ja → Zh are the same data that included in the MMT baseline training data.

3.2 LLM Systems

We use LLaMA2-7B as the backbone because it is permitted for the constraint track of WMT24. We reuse the framework of ALMA (Xu et al., 2023) to conduct fine-tuning, however, we discard their first stage of monolingual continue training.

We set the training batch as 32 and accumulated 4-step gradients. The learning rate is set as 2e-5. The model was trained for one epoch using bf16 precision. The beam size is set as 5 for inference.

For the fine-tuning dataset, we further apply the quality estimation method described in Section 3.1 to filter out data with a QE score below a certain threshold. Then, we sample a certain number of bitext from the filter dataset. For example, in Table 4, the number of samples with a score above 89 is 53k, all of which are used for fine-tuning. Additionally, we sample data with scores higher than 87 at various levels, such as 53k, 100k, and 300k. We fine-tune LLaMA2 with different kinds of data to show the impact of data qualities.

3.3 NMT+LLM Systems

Previous studies have shown that leveraging Large Language Models (LLMs) to post-edit the outputs of supervised Neural Machine Translation (NMT) models can reduce translationese and enhance translation quality (Chen et al., 2023). This strategy has proven effective with LLMs such as ChatGPT (Chen et al., 2023), GPT-4 (Raunak et al., 2023), PaLM (Xu et al., 2024b), and LLaMA-2 (Ki and Carpuat, 2024). Specifically, post-editing utilizes LLMs either to refine the outputs of supervised NMT models or to perform "Self-Refinement" on their own outputs. Furthermore, Ki and Carpuat (2024) demonstrate that tuning LLMs with error-annotated translations can further enhance performance.

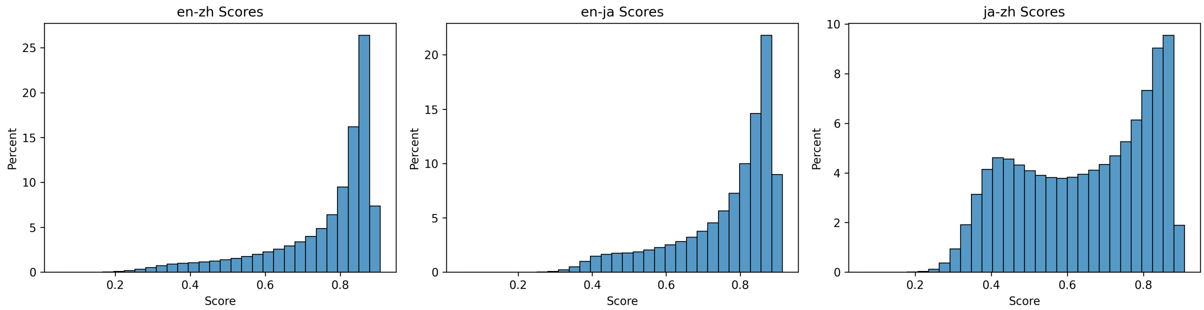

Figure 1: Comet score distributions for WMT-24 constraint training data on en→zh, en→ja, and ja→zh directions.

In this paper, we explore the effectiveness of Post-Editing (PED) in improving translation quality for the English-to-Chinese direction. We focus on a training-free PED approach due to computational constraints, utilizing pre-trained open LLMs to edit the outputs of our supervised NMT models. Given the limited Chinese capability of LLaMA2, we employ Tower-LLMs (Alves et al., 2024) (Tower-Instruct 7B and 13B), which have been continuously pre-trained on monolingual corpora including Chinese. Additionally, we implement a Quality Estimation-guided Post-Editing (QE-based PED) approach, where the NMT outputs and post-edited outputs are selected based on the higher QE score using COMETKiwi (wmt22-cometkiwi-da).

4 Experimental Setups

4.1 Systems

In this section, we briefly describe the systems we implemented. It is important to note that some of our implementations were focused only on the English-to-Chinese direction, specifically for FT-LLaMA2, Tower-Instruct, the PED system, and the QE-based PED system.

mT-large. A multilingual Transformer-large model trained in many-to-many directions using the "Cleaned" data (see Table 1 and Section 3.1 for details). It consists of 12 layers with 16 attention heads, d = 1,024, and d_{ff} = 4,096.

mT-large + QE. This model shares the same architecture and hyper-parameter settings as the *mT-large* model but is trained using the "QE-filtered" data outlined in Table 1.

mT-large + QE + FT. The *mT-large + QE* model was further fine-tuned on direction-specific data.

FT-LLaMA2. We use supervised fine-tuning to fine-tune LLaMA2. Detailed settings can be found in Section 3.2.

Tower-Instruct. We directly evaluate the performance of the Tower-Instruct models for comparison with our systems.

Self-Refined PED. We prompt the Tower-Instruct model to post-edit the translations they originally generated.

PED system. We prompt Tower-Instruct models to post-edit the outputs generated by our supervised NMT system (*mT-large + QE + FT*).

QE-guided PED system. We determined the final outputs by selecting between the NMT outputs and the post-edited outputs, based on the higher QE score as determined by COMETKiwi.

4.2 Data

For training, we utilize both the "Cleaned" and "QE-filtered" datasets, see details in section 2. For evaluation, we employ previous WMT validation and test sets as our validation set, and Flores, Ntrex as our test set.

4.3 Implementation and Evaluation

For our Neural Machine Translation (NMT) systems, we utilize the Fairseq toolkit (Ott et al., 2019) for both training and inference. For Large Language Model systems, we employ the Transformers toolkit for training and inference. To evaluate our models, we report detokenized SacreBLEU[4], ChrF++(Popović, 2017), and COMET (Rei et al., 2020) (wmt22-comet-da) scores.

[4] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

ID	Methods	#Param	FLORES-Devtest			NTREX			WMT-24 Official	
			BLEU	ChrF	COMET	BLEU	ChrF	COMET	BLEU	ChrF
					English→Japanese (NMT Systems Only)					
①	mT-large	419M	35.9	39.0	89.48	26.9	33.8	85.79	29.8	26.2
②	① + QE	419M	36.6	39.8	90.00	27.3	34.5	86.95	34.2	29.6
③	② + FT	419M	37.1	40.3	90.24	28.3	35.1	87.18	34.7	30.1
					Japanese→Chinese (NMT Systems Only)					
④	mT-large	419M	33.9	29.2	86.64	27.5	24.9	82.26	22.5	21.6
⑤	④ + QE	419M	34.0	29.1	87.04	27.6	25.0	82.77	22.7	22.0
⑥	⑤ + FT	419M	34.0	29.1	87.00	27.8	25.0	82.53	22.9	21.6

Table 2: Translation quality on NTREX, FLORES, and WMT test sets for the English→Japanese and Japanese→Chinese directions. 'FT' denotes directional Fine-Tuning, and 'QE' represents using QE-filtered training data. We use percentage for COMET scores.

ID	Methods	#Param	FLORES-Devtest			NTREX			WMT-24 Official	
			BLEU	ChrF	COMET	BLEU	ChrF	COMET	BLEU	ChrF
					NMT Systems					
①	mT-large	419M	42.2	35.0	84.94	33.3	28.7	79.20	-	-
②	① + QE	419M	43.8	36.0	86.21	34.7	29.7	81.21	-	-
③	② + QE + FT	419M	43.9	36.2	86.12	35.0	29.7	80.95	33.5	31.6
					LLM Systems					
④	FT-LLama2	7B	34.6	31.2	86.60	-	-	-	-	-
⑤	Tower-Instruct	7B	42.3	37.4	88.09	35.2	31.1	85.42	36.2	33.2
⑥	Tower-Instruct	13B	43.2	38.0	88.12	36.2	32.0	85.36	38.5	35.3
					NMT + LLM Systems					
⑦	Self-Refined PED (⑤)	7B	40.3	36.1	85.61	34.1	30.4	83.79	36.0	33.0
⑧	PED (③ + ⑤)	7.42B	39.7	35.8	83.68	31.3	28.3	78.80	38.1	34.9
⑨	QE-based PED (③ + ⑤)	7.42B	40.7	36.1	86.22	32.5	29.2	81.40	38.2	35.3

Table 3: Translation quality on NTREX, FLORES-200, and WMT-24 test sets for the English→Chinese direction. For WMT-24, we report BLEU and ChrF scores as returned by the OCELoT submission system.

5 Results and Analyses

In this section, we present the final results of our experiments and discuss the findings. Table 3 and 2 show the results of English→Chinese and the other two directions (en→ja and ja→zh) on Flores-devtest, Ntrex, and WMT24 official test sets.

5.1 Quality-Estimation Filtering improves NMT systems

Our key finding is that implementing Quality-Estimation (QE) Filtering effectively reduces low-quality data samples, leading to improved NMT system performance. Specifically, we observed BLEU score improvements of +4.4 and +0.2 for the English→Japanese and Japanese→Chinese directions, respectively, on the WMT24 official test sets. For the English→Chinese direction, we observed BLEU gains of +1.6 on the Flores-devtest and +1.4 on the Ntrex test sets. Similar positive performance improvements were also noted across other metrics, such as ChrF and COMET. These results indicate that filtering training samples with low COMET scores enables our supervised NMT system to generate higher-quality translations.

5.2 Fine-tuned LLaMA2 and Data Quality

We conduct experiments on LLaMA2-7B in English to Chinese translation direction, where we collect 300K parallel samples from the training set, controlling the QE scores are all higher than 87. In Table 3, ④ shows the results. It is easy to see that the fine-tuned LLaMA2 results in the best COMET performance (86.60) on the Flores benchmark. However, the results on surface-level metrics, such as BLEU and ChrF, significantly lag

Language	Data	BLEU	COMET
LLama2-7B	10k (Cleaned)	28.0	82.7
LLama2-7B	100k (Cleaned)	35.7	85.6
LLama2-7B	53k (COMET > 87)	**36.1**	86.1
LLama2-7B	53k (COMET > 89)	33.5	84.3
LLama2-7B	100k (COMET > 87)	35.5	85.7
LLama2-7B	300k (COMET > 87)	34.6	**86.6**

Table 4: Evaluation results of fine-tuned LLama2-7B models for the English→Chinese direction on the Flores-devtest set. 'Cleaned' indicates random sampling from the 'Cleaned' training dataset, while 'COMET>x' refers to the sampling of data with COMET scores greater than x.

behind encoder-decoder-based NMT systems by 7.6 and 3.8 points, respectively.

We further control the fine-tuning data quality to show the impact. We select 10K and 100K samples from the cleaned dataset (See Table 1). To further improve the quality of parallel semantic alignment, we score all of the 39M cleaned training samples using COMET, and then we construct fine-tuning sets under the following settings:

- We selected all 53k samples with very high COMET scores, using a threshold of 89.
- We then lowered the score threshold to 87 and selected another 53k samples.
- We extend the number of samples with scores higher than 87 to 100k and 300k.

Table 4 shows the corresponding results after fine-tuning using datasets with different qualities. We observe that: 1) Simply extending the fine-tuning set from 10k to 100k largely improves the resulting performance. 2) However, no clear improvements can be observed when further raising the fine-tuning data QE quality. E.g., using 100k trivial samples (after data cleaning, QE score lower than 80) achieves comparable performance to that of using 100k samples with a QE score higher than 87. Additionally, fine-tuning with samples that have extremely high QE scores (COMET > 89) even resulted in a decline in translation quality compared to using 53k samples with relatively lower QE scores (COMET > 87). 3) Further extending the fine-tuning size from 100k to 300k yields no clear improvements.

Our experiments suggest that simply enhancing the quality of fine-tuning data for LLMs, at least when using COMET as the central measure of quality, is not a promising approach.

5.3 Post-Editing Enhances Translation Quality

As shown in Table 3, using the Tower-Instruct 7B LLM to post-edit the outputs of our strongest supervised NMT model (PED (③ + ⑤)) resulted in large improvements, with BLEU and ChrF gains of +4.6 and +3.3, respectively, over the NMT model alone on the WMT24 official test set. Notably, this post-editing approach also outperformed direct translation with Tower-Instruct 7B, achieving additional gains of +1.9 BLEU and +1.7 ChrF. In contrast, applying the Tower-Instruct model to post-edit its own generated translations (self-refined PED) resulted in negative improvements across all test sets. These findings suggest that integrating supervised NMT models with LLMs is a promising strategy for enhancing translation quality by leveraging the strengths of both systems.

Furthermore, Table 3 demonstrates that the QE-guided PED system (QE-based PED (③ + ⑤)) can further improve translation quality, as evidenced by the positive performance gains across the Flores-devtest, Ntrex, and WMT24 official test sets. In particular, the QE-guided PED system, utilizing Tower-Instruct 7B as the LLM backbone, achieved performance on par with Tower-Instruct 13B in the ChrF metric on the WMT24 official test set.

Despite the promising results on the WMT-24 Official test set, we found this Post-Editing approach delivered negative performance improvements on Flores and Ntrex sets (Table 3).

6 Conclusions

In this paper, we investigate three aspects of using LLMs for translation: 1) Comparison with Encoder-Decoder NMT Systems: directly fine-tuning LLaMA2 on a relatively small bitext dataset (100K) yields COMET results comparable to those of strong encoder-decoder NMT systems trained on over 50 million parallel sentence pairs. However, this approach significantly underperforms in surface-level metrics such as BLEU and ChrF. 2) Impact of Data Quality: properly filtering samples with low COMET scores largely improves encoder-decoder systems, however, no clear improvements can be observed for LLMs when further controlling the fine-tuning set with higher COMET scores. 3) Combining NMT Systems with LLMs: lastly, we show that combining NMT systems with LLMs via post-editing generally yields the best performance in our experiments.

Acknowledgements

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080.

References

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and André FT Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. Crosssum: Beyond english-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*.

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*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. Ccaligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

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*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83.

Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2023. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Viet Lai, Chien Nguyen, Nghia Ngo, Thut Nguyn, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. *arXiv preprint arXiv:2405.12915*.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using qe metrics for mt data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Surangika Ranathunga, Nisansa De Silva, Velayuthan Menan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

David Stap and Ali Araabi. 2023. Chatgpt is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing llm abilities. *arXiv preprint arXiv:2405.20089*.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. *arXiv preprint arXiv:2404.11201*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, et al. 2023b. 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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11.

Di Wu, Shaomu Tan, David Stap, Ali Araabi, and Christof Monz. 2023. Uva-mt's participation in the wmt 2023 general translation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 175–180.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024b. Llmrefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

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*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.