IKUN for WMT24 General MT Task: LLMs Are here for Multilingual Machine Translation

Baohao Liao[1,2] **Christian Herold**[2] **Shahram Khadivi**[2] **Christof Monz**[1]
[1]Language Technology Lab, University of Amsterdam
[2]eBay Inc., Aachen, Germany
b.liao@uva.nl

Abstract

This paper introduces two multilingual systems, *IKUN* and *IKUN-C*, developed for the general machine translation task in WMT24. IKUN and IKUN-C represent an *open system* and a *constrained system*, respectively, built on Llama-3-8b and Mistral-7B-v0.3. Both systems are designed to handle all 11 language directions using a single model. According to automatic evaluation metrics, **IKUN-C achieved 6 first-place and 3 second-place finishes among all constrained systems, while IKUN secured 1 first-place and 2 second-place finishes across both open and constrained systems**. These encouraging results suggest that large language models (LLMs) are nearing the level of proficiency required for effective multilingual machine translation. The systems are based on a two-stage approach: first, continuous pre-training on monolingual data in 10 languages, followed by fine-tuning on high-quality parallel data for 11 language directions. The primary difference between IKUN and IKUN-C lies in their monolingual pre-training strategy. IKUN-C is pre-trained using constrained monolingual data, whereas IKUN leverages monolingual data from the OSCAR dataset. In the second phase, both systems are fine-tuned on parallel data sourced from NTREX, Flores, and WMT16-23 for all 11 language pairs.[1]

1 Introduction

Large language models (LLMs) (Touvron et al., 2023; Dubey et al., 2024; Jiang et al., 2023; OpenAI, 2023) serve as a crucial foundation for a wide range of applications. One significant advantage of LLMs is their ability to be applied across various tasks, thereby simplifying deployment processes. However, the application of LLMs to multilingual machine translation (MT) presents several challenges:

- Most LLMs are pre-trained on one or a few dominant languages, making direct fine-tuning on multilingual data insufficient for ensuring optimal performance, particularly for low-resource languages, which are often underrepresented in the training data.

- It remains unclear whether these LLMs, primarily pre-trained on a limited number of languages, effectively facilitate transfer learning across different languages (Tan et al., 2024).

- The large-scale nature of most LLMs presents significant challenges for efficient fine-tuning, particularly for researchers and practitioners with limited computational resources.

In the WMT24 general MT task (Kocmi et al., 2024a), our objective is to assess the capability of LLMs for multilingual MT, as an alternative to training bilingual systems from scratch (Wu et al., 2023). This paper provides a detailed account of how we developed our final multilingual system using LLMs for both the constrained and open tracks.

Firstly, we identified that certain LLMs exhibit inefficiencies in tokenizing sentences from languages that are underrepresented in the pre-training data. To address this, we extended the existing vocabulary to reduce the tokenized sentence length, thereby enhancing training efficiency. Secondly, we enriched the LLMs with knowledge across the 10 target languages through continued pre-training. This step is particularly crucial for underrepresented languages, as it facilitates transfer learning. Finally, we fine-tuned the models using high-quality parallel datasets across all 11 pairs.

Through this streamlined approach, our constrained multilingual system, IKUN-C, secured 6 first-place and 3 second-place rankings in the automatic evaluation. Our open multilingual system, IKUN, achieved 1 first-place and 2 second-place rankings across the open and constrained tracks.

[1]Please read our newest verion at https://arxiv.org/abs/2408.11512

Proceedings of the Ninth Conference on Machine Translation, pages 263–269
November 15-16, 2024

These encouraging results demonstrate that LLMs can be effectively adapted for multilingual MT, broadening access to speakers of diverse languages.

2 Pre-trained LLM

LLMs are pre-trained on extensive web-scale data, encompassing a vast repository of general knowledge applicable to various tasks. Previous studies (Xu et al., 2024a,b) have demonstrated that LLMs can substantially enhance the performance of multilingual MT. Building on this insight, we adopt a pre-trained LLM as the foundation for our system.

IKUN is an open system developed with meticulous consideration of available resources and system capabilities. For this purpose, we selected Llama-3 (Dubey et al., 2024), one of the most advanced open-source LLMs available at the time of this competition. Due to constraints on computational resources, we opted for the 8B version[2] instead of the 70B version. A significant factor in our choice of Llama-3 was its strong support for multilingual applications, as evidenced by the efficiency with which its tokenizer handles all 11 languages involved in this competition (See §3). We also tried the instruct version, but it is worse than the pre-trained version.

IKUN-C is a constrained system based on Mistral-7B-v0.3 (Jiang et al., 2023), one of the three LLMs permitted for the constrained track. Prior to selecting Mistral-7B-v0.3[3], we conducted continuous pre-training on all three allowed LLMs — namely, Llama-2-7B, Llama-2-13B (Touvron et al., 2023), and Mistral-7B — using a subset of our monolingual data (approximately 1B tokens). The pre-training loss demonstrated that Mistral-7B outperformed Llama-2-7B and performed comparably to Llama-2-13B, leading us to select it as our architecture of choice.

3 Tokenizer Efficiency

A significant challenge in applying LLMs to multilingual MT lies in the efficiency of their tokenizers. These models are typically pre-trained on one or a few dominant languages, and when their tokenizers are applied to low-resource languages, they produce disproportionately long sequences of sub-words. This inefficiency leads to excessive GPU memory consumption during training.

To evaluate the efficiency of the tokenizer, we focus on comparing the length differences between tokenized English sentences and their corresponding non-English counterparts. Specifically, we define the length ratio as:

$$\text{length ratio} = \frac{\text{len}(\text{tokenizer}(x))}{\text{len}(\text{tokenizer}(y))}$$

where y represents the English sentence, and x denotes the paired non-English sentence. A smaller length ratio (close to 1) is desired, since it means that the tokenizer can encode the non-English sentence as efficient as the English sentence.

To facilitate a comparison of length ratios across different languages, English-centric multilingual data is essential. Fortunately, the FLoRes-200 dataset (Costa-jussà et al., 2022) possesses this characteristic. In the devtest and test sets of FLoRes-200, every English sentence is paired with translations in various other languages. We concatenate all sentences from the devtest set and compute the length ratio for each language, as illustrated in Figure 1. We also include NLLB's tokenizer (Costa-jussà et al., 2022) for a comparison.

We can observe: (1) NLLB consistently exhibits the smallest length ratio across all languages, likely due to its extensive optimization for hundreds of languages, thus serving as a lower bound in this context. (2) Mistral-v0.3 and Llama-3 demonstrate a notably high length ratio for Hindi, suggesting that Hindi is underrepresented in the pre-training data. (3) Compared to NLLB, the tokenizer of Mistral-v0.3 is significantly less efficient for Chinese, Japanese, Hindi, and Icelandic.

We opted to expand the vocabulary by incorporating new sub-words to reduce the length of tokenized sentences, thereby enhancing training efficiency. However, this approach introduces a trade-off between the addition of new sub-words and training performance. The embeddings for the newly introduced sub-words are initially untrained, and a substantial increase in sub-words may necessitate additional iterations of continuous pre-training. Consequently, our strategy for adding sub-words prioritizes those from languages with higher length ratios.

For our open system, IKUN, we didn't modify its tokenizer, since Llama-3 tokenizer is already efficient enough, only except for Hindi. For our constrained system, IKUN-C, we expanded its vocabulary for Chinese, Japanese, Hindi, and Icelandic through the following steps: (1) Generate

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B
[3] https://huggingface.co/mistralai/Mistral-7B-v0.3

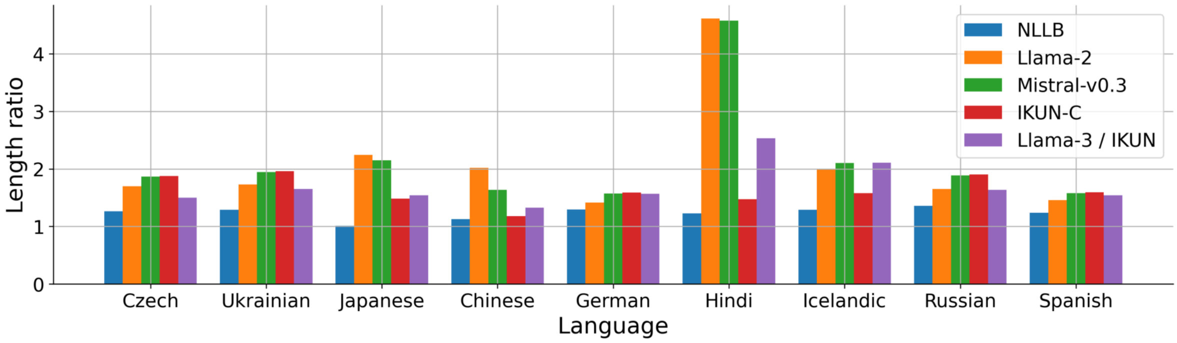

Figure 1: Tokenizer efficiency for various LLMs and languages. The larger the length ratio is, the less efficient the tokenizer is. We add new sub-words, from Chinese, Japanese, Hindi and Icelandic, to the Mistral-v0.3 vocabulary to construct the IKUN-C vocabulary. IKUN uses the Llama-3 tokenizer without any modification.

Language pair	**Num.**	**Language pair**	**Num.**
cs-uk	8768	uk-cs	8768
ja-zh	12858	zh-ja	12858
en-zh	36647	zh-en	34650
en-cs	30120	cs-en	28123
en-de	35564	de-en	33567
en-hi	11020	hi-en	9023
en-is	8010	is-en	6013
en-ja	18113	ja-en	16116
en-ru	32840	ru-en	30843
en-es	13808	es-en	11811
en-uk	11961	uk-en	9964
en-fr	4006	fr-en	4006
Total:		429457	

Table 1: Number of parallel sentences.

a new vocabulary of 12K sub-words using monolingual data from the past two years for these four languages from News Crawl[4]; and (2) Merge the new vocabulary with the original. The efficiency of the resulted IKUN-C tokenizer is shown in Figure 1, demonstrating more efficiency for these four languages, especially for Hindi.

4 Experiments

We mainly follow the pipeline from Xu et al. (2024a), i.e. continuous pre-training on monolingual data for all 10 languages, and followed by fine-tuning on parallel data for all 11 language pairs.

4.1 Continuous Pre-training

Given that our selected LLMs, specifically Llama-3-8B and Mistral-7B-v0.3, are primarily pre-trained on English, it is necessary to incorporate knowledge from other languages through further pre-training, with particular emphasis on low-resource languages. Additionally, the word embeddings for the newly introduced sub-words in IKUN-C must also undergo training.

For the open system IKUN, we utilize monolingual data from the Oscar dataset (Suárez et al., 2020), covering all 10 target languages. We adopt the sampling strategy outlined by Xu et al. (2024a), described as:

$$P(l) \propto (\frac{D_l}{\sum_{l' \in L} D_{l'}})^{\frac{1}{T}} \quad s.t. \quad \sum_{l' \in L} P(l') = \frac{9}{10}$$

where D_l represents the number of words in language l[5], T is the temperature parameter (set to 6), and L denotes the set of all languages except English. The sampling probability for English is fixed at 1/10. The experimental settings for continuous pre-training are detailed in Table 2. We approximately pre-trained IKUN on an additional 8B tokens.

In the constrained system, only the provided data sources are permitted for use[6]. The IKUN-C system utilizes monolingual data from the News Crawl dataset for 9 languages, with the exception of Spanish, as the use of Spanish monolingual data from News Crawl is restricted. For Spanish, we incorporate monolingual data from the Leipzig Corpora (Goldhahn et al., 2012). Additionally, for Hindi, we augment the dataset with monolingual data from News Commentary due to the relatively limited amount of Hindi data available in the News Crawl dataset. This adjustment is crucial because Hindi is underrepresented in the pre-training of Mistral-7B-v0.3. Our experimental settings closely align with those detailed in Table 2.

[4]https://data.statmt.org/news-crawl/

[5]https://huggingface.co/datasets/oscar-corpus/OSCAR-2301

[6]https://www2.statmt.org/wmt24/mtdata/

Hyper-parameter	Continuous pre-training	Finetuning
sampling probability	cs,de,en,es,hi,is,ja,ru,uk,zh = 0.1,0.13,0.1,0.13,0.08,0.05,0.08,0.13,0.08,0.12	
duration	60K steps	1 epoch
batch size	64	128
sequence length	2048	max source length=512, max target length=512
learning rate (lr)	2e-5	2e-4
warmup ratio	0	0.01
weight decay	0.01	0.01
lr scheduler	cosine	inverse_sqrt
training type	full finetuning	LoRA $r = 64$ for all layers

Table 2: Experimental setting for continuous pre-training and subsequent finetuning.

System	Metric	cs-uk	en-cs	en-de	en-es	en-hi	en-is	en-ja	en-ru	en-uk	en-zh	ja-zh
IKUN-C	MetricX ↓	2.4	4.3	2.0	3.5	7.1	4.9	4.3	4.7	4.7	4.2	6.2
	CometKiwi ↑	0.648	0.618	0.641	0.666	0.499	0.657	0.669	0.649	0.622	0.624	0.519
	AutoRank ↓	3.0	4.7	3.8	3.4	5.5	3.7	3.9	3.9	3.9	3.5	5.5
	Place in constrained ↓	2	5	1	1	1	1	3	1	1	2	2
IKUN	MetricX ↓	1.6	3.7	1.8	3.3	9.4	4.3	3.7	4.1	3.7	4.0	5.4
	CometKiwi ↑	0.664	0.638	0.668	0.687	0.428	0.666	0.696	0.675	0.661	0.646	0.544
	AutoRank ↓	2.3	3.9	3.0	2.8	7.7	3.2	3.1	3.2	2.8	3.1	4.4
	Place in constrained&open ↓	2	5	4	3	5	1	6	3	2	5	6

Table 3: Preliminary results of our systems on the WMT24 test sets, taken from Kocmi et al. (2024b). The final human evaluation results are not released yet. "-" here means →. I.e. cs-uk is cs→uk.

4.2 Subsequent Fine-tuning

Previous studies (Wu et al., 2024; Zhou et al., 2023) have demonstrated that the quality of fine-tuning data is a critical factor in achieving optimal performance. Liao et al. (2021) further indicates that increasing the amount of back-translation data does not necessarily lead to better outcomes. In light of this, we exclusively utilize high-quality parallel data for the fine-tuning phase.

The high-quality parallel data is primarily sourced from FloRes-200 (Costa-jussà et al., 2022), NTREX-128 (Federmann et al., 2022), and previous WMT16-23 general MT/news tasks (Kocmi et al., 2023, 2022; Akhbardeh et al., 2021; Barrault et al., 2020, 2019, 2018; Bojar et al., 2017, 2016).

FloRes-200: As the FloRes-200 dataset provides parallel sentences across multiple languages, we leverage all 11 translation directions from the devtest and test sets. Importantly, our fine-tuning approach is not limited to the required directions listed in Table 3; instead, we fine-tune the model on both translation directions, e.g., en↔de, to facilitate broader applicability of the final model.

NTREX-128: We also incorporate parallel sentences from NTREX-128 for from-English translation directions, i.e. en→XX. In accordance with Federmann et al. (2022), which recommends using the en→XX translation direction, our fine-tuning is confined to these directions rather than adopting a bidirectional approach. An exception is made for the en-fr pair, where bidirectional fine-tuning is applied due to the limited availability of parallel data for this pair (absent in previous WMTs).

Past WMTs: Additionally, we extract parallel sentences from the development and test sets of the WMT16-23 general MT tasks, provided they contain the necessary translation directions. For these sentences, we employ a bidirectional fine-tuning strategy.

The statistics for all parallel sentences are presented in Table 1. Notably, all systems are fine-tuned at the sentence level. Given that WMT development and test sets are at the document level, models could alternatively be fine-tuned at the document level or reformatted into a conversational structure for fine-tuning. This latter approach might be more effective for context-aware translation, as WMT24 applies context-based human evaluations. We reserve this exploration for future work. The fine-tuning setting is listed in Table 2.

4.3 Results

We present the preliminary results reported by Kocmi et al. (2024b) in Table 3, which includes four evaluation metrics. Both MetricX-23-XL (Juraska et al., 2023) and CometKiwi-DA-XL (Rei et al., 2023) have demonstrated strong correlations with human evaluation (Freitag et al., 2023). AutoRank (Kocmi et al., 2024b), a normalized composite metric derived from MetricX and CometKiwi, scales the scores of each metric linearly to span the range from 1 to the total number of systems in a given language pair. The final automatic ranking is obtained by averaging these normalized scores. AutoRank can thus be considered a measure of the overall rank across all systems and tracks. Additionally, we report the rankings of our systems across various tracks.

It is noteworthy that both of our systems are multilingual, designed to handle all language pairs. The IKUN-C system, in particular, demonstrates promising performance in the constrained track, achieving 6 first-place and 3 second-place finishes. In both the open and constrained tracks, IKUN maintains strong performance, securing 1 first-place and 2 second-place positions, even when compared to systems that may leverage additional open-source data or specialize in a limited set of language pairs.

5 Conclusion

In this paper, we present a methodology for effectively adapting pre-trained LLMs to the task of multilingual machine translation. Our approach involves three primary steps: (1) expanding the vocabulary to accommodate languages that are underrepresented in the pre-training data, when necessary; (2) continuing pre-training the LLM on monolingual data to enhance its knowledge of underrepresented languages and to train the embeddings of newly introduced sub-words; and (3) fine-tuning the LLM on high-quality parallel data. Our experimental results demonstrate the efficacy of this straightforward pipeline, with IKUN-C securing 6 first-place finishes in the constrained track, and IKUN achieving 1 first-place ranking in both the open and constrained tracks.

Acknowledgments

We thank eBay Inc. for the computation support. This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondrej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussà, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 1–88. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubesic, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 1–55. Association for Computational Linguistics.

Loïc Barrault, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 1–61. Association for Computational Linguistics.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 304–323. Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 169–214. Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 131–198. The Association for Computer Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, and et al. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and et al. 2024. The llama 3 herd of models.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George F. Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 578–628. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 759–765. European Language Resources Association (ELRA).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, and et al. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task:
the LLM era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024b. Preliminary wmt24 ranking of general mt systems and llms.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popovic, and Mariya Shmatova. 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, WMT 2023, Singapore, December 6-7, 2023*, pages 1–42. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popovic. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1–45. Association for Computational Linguistics.

Baohao Liao, Shahram Khadivi, and Sanjika Hewavitharana. 2021. Back-translation for large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 418–424. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Ricardo Rei, Nuno Miguel Guerreiro, José Pombal, Daan van Stigt, Marcos V. Treviso, Luísa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 841–848. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1703–1714. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking overall zero-shot multilingual translation via tiny multi-parallel data. *CoRR*, abs/2401.12413.

Di Wu, Shaomu Tan, David Stap, Ali Araabi, and Christof Monz. 2023. Uva-mt's participation in the WMT23 general translation shared task. *CoRR*, abs/2310.09946.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *CoRR*, abs/2401.08417.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.