TOWER-V2: Unbabel-IST 2024 Submission for the General MT Shared Task

Ricardo Rei[*1] , **José Pombal**[*1,2,4] , **Nuno M. Guerreiro**[*1,2,4,5] , **João Alves**[*1] , **Pedro H. Martins**[*1]
Patrick Fernandes[2,3,4] , **Helena Wu**[1] , **Tania Vaz**[1] , **Duarte M. Alves**[2,4] , **Amin Farajian**[1]
Sweta Agrawal[2] , **Antonio Farinhas**[2,4] , **José G.C. de Souza**[1], **André F. T. Martins**[1,2,4]
[1]Unbabel
[2]Instituto de Telecomunicações [3]Carnegie Mellon University
[4]Instituto Superior Técnico & Universidade de Lisboa (Lisbon ELLIS Unit)
[5]MICS, CentraleSupélec, Université Paris-Saclay

Abstract

In this work, we present TOWER-V2, an improved iteration of the state-of-the-art open-weight TOWER models, and the backbone of our submission to the WMT24 General Translation shared task. TOWER-V2 introduces key improvements including expanded language coverage, enhanced data quality, and increased model capacity up to 70B parameters. Our final submission combines these advancements with quality-aware decoding strategies, selecting translations based on multiple translation quality signals. The resulting system demonstrates significant improvement over previous versions, outperforming closed commercial systems like GPT-4O, CLAUDE-SONNET-3.5, and DEEPL even at a smaller 7B scale.

1 Introduction

Large Language Models (LLMs) are making strides towards becoming the *de facto* solution for multilingual machine translation (MMT). Many works have shown that it is possible to adapt LLMs for translation and achieve state-of-the-art results (Zhang et al., 2023; Wei et al., 2023; Alves et al., 2023; Reinauer et al., 2023; Zhu et al., 2024).

One such example is our recent work on TOWER (Alves et al., 2024), which demonstrates that open NMT models like NLLB200 can be outperformed by adapting an LLM to translation. Specifically, we continue the pre-training of LLaMA-2 (Touvron et al., 2023) on both monolingual and parallel data, and fine-tune the resulting model on high-quality instructions covering several MT-related tasks. This approach requires much less parallel training data than traditional NMT and preserves the general capabilities of the LLM to respond to various prompts.

For the WMT24 General Translation task (Kocmi et al., 2024a), we enhance TOWER by significantly improving its training data, by extending its language support from 10 to 15 languages — including low-resource ones like Icelandic —, and by scaling the underlying model to 70 billion parameters. Furthermore, because the WMT24 General Translation task focuses on paragraph-level translation instead of sentence-level, we also experiment with full-document translation and longer contexts, where TOWER originally struggled. These key improvements result in TOWER-V2 7B and 70B.

For our primary submission, we combine TOWER-V2 70B with Quality-Aware Decoding (QAD) strategies (Fernandes et al., 2022), such as Minimum Bayes Risk decoding (MBR) and Tuned Reranking (TRR). These techniques use reward models during inference to select the best candidate from a set of generated samples, enhancing the overall output quality.

We report our results, including the human evaluation and final submission, in Section 5. By outperforming strong commercial systems like GPT-4, CLAUDE-SONNET-3.5, and DEEPL across the board, TOWER-V2 — even at 7B parameters — challenges the belief that in MMT there must be a trade-off in performance between high- and low-resource language pairs (Fernandes et al., 2023).

Our contributions are:

- We show that expanding from 10 to 15 languages maintains the quality of translations for the initial 10 and significantly improves the newly added languages.
- We significantly improve the paragraph- and document-level translation capabilities of the previous TOWER.
- We demonstrate that scaling the model from 7 to 70B parameters yields improvements, indicating that increased capacity benefits not only general LLM abilities but also task-specific performance.

[*]Core Contributor. ✉ ai-research@unbabel.com

Proceedings of the Ninth Conference on Machine Translation, pages 185–204
November 15-16, 2024 ©2024 Association for Computational Linguistics

- We analyze the impact of QAD on larger models than those studied by Fernandes et al. (2022), showing that MBR decoding outperforms TRR according to both automatic metrics and human evaluation.

2 Overview of the Shared Task

The primary aim of the general machine translation shared task is to evaluate the ability of various models to translate across different domains, genres, and possibly modalities (e.g., speech). This year's shared task, compared to previous editions, emphasizes English→X (en→xx) and Non-English→Non-English (xx→yy) language pairs.[1]

The WMT24 test sets include source sentences from four domains: news articles, social media posts, speech (machine-generated transcripts), and literary texts. Additionally, all test sets from this year are focusing on the paragraph level rather than sentence-level.

Throughout this paper we will evaluate several of our models using both automatic and human evaluation; yet, for the shared task only primary submissions are evaluated, and final results are based solely on human evaluation using the ESA protocol (Kocmi et al., 2024c).

3 TOWER-V2: A New Translation LLM

We create TOWER-V2 by improving upon the original TOWER recipe: continued pre-training of a base model on a multilingual dataset with billions of tokens and subsequent supervised fine-tuning for translation-related tasks.

We focus on three key areas: 1) careful refinement of the training data; 2) expansion of language coverage to support all of the shared task's languages; 3) scaling up model capacity.

Improving the training data. To enhance the general translation capabilities of TOWER, we mainly focus on improving the quality of its training data, be it for translation, post-translation, or general instructions.

For continued pre-training (CPT), we train on monolingual data from sources of superior quality, and apply more aggressive quality and length filters on the parallel data.

Regarding the supervised fine-tuning (SFT) phase, we use data created by humans — similarly to the previous version of TOWER— and introduce high-quality synthetic data. Human translations are sourced from well-known translation benchmarks. We go beyond simple sentence-level translation by transforming sentence-level to document-level data or into multi-parallel translation data (translating a single source sentence into multiple languages). When language variants are available, we include them in the training prompt (e.g. Chinese (simplified) vs Chinese (Taiwan)). All datasets were carefully filtered[2] and converted to instructions using a diverse set of templates.

Improving post-translation data and general instructions. Data from tasks like APE, MQM evaluation, and translation ranking are carefully filtered using several quality signals. Similarly to XTOWER (Treviso et al., 2024), APE and MQM evaluation always expect the model to return a "translation correction," so we always ensure that the post-edition (PE) is deemed better than the original translation according to several metrics. For translation ranking, we choose only samples where there is significant alignment between human annotations and automatic metrics.

Like in the previous TOWER version, we aim to build a model that adheres to different prompts and can work as a general multilingual LLM. Thus, we include filtered and adapted multilingual general-purpose instruction data from publicly available high quality datasets such as AYA (Singh et al., 2024).

Going from 10 to 15 Languages. We extend the language support of TOWER-V2 to Czech, Icelandic, Hindi, Ukrainian, and Japanese by adding training data of these languages to both CPT and SFT stages. For CPT, we add monolingual and parallel training data, increasing the total number of training tokens considerably. Aside from to-/from-English language pairs, we also include Czech-Ukrainian and Japanese-Chinese (and vice-versa) parallel data. In the SFT stage, we mostly add translation data for the new language pairs.

More Paragraphs/Documents. In addition to the sentence-level parallel data we also add parallel documents to the CPT stage. For SFT, we sample high quality monolingual documents and per-

[1] The complete list of language pairs for this year's task includes: Czech→Ukrainian, Japanese→Chinese, and English→Chinese, Czech, German, Hindi, Icelandic, Japanese, Russian, Spanish (Latin America), Ukrainian

[2] We found low-quality translations even on datasets built by professionals.

	WMT24										
Model	en→de	en→es	en→cs	en→ru	en→uk	en→is	en→ja	en→zh	en→hi	cs→uk	ja→zh
Baselines											
NLLB-54B	7.23 9	7.05 9	8.63 9	7.51 9	8.42 8	9.66 9	5.46 8	10.18 8	4.31 6	4.16 7	11.33 9
GPT-4o	1.41 6	1.57 7	1.48 6	1.39 6	1.42 6	2.31 7	1.04 5	1.65 5	1.19 4	0.94 4	3.42 6
Claude-Sonnet-3.5	1.33 5	1.52 6	1.34 5	1.27 5	1.30 5	2.19 6	0.95 4	1.53 4	1.14 3	0.86 3	3.11 4
DeepL	1.81 8	2.10 9	1.71 7	2.21 8	1.44 6	—	3.95 7	2.22 7	—	1.40 5	7.36 9
Tower											
Tower-v1 13B	1.61 7	1.67 8	—	1.64 7	—	—	—	1.82 6	—	—	—
Tower-v2 7B	1.41 6	1.42 5	1.39 5	1.41 6	1.36 5	1.90 5	1.10 5	1.71 5	1.57 5	0.82 3	3.66 7
Tower-v2 70B	1.26 4	1.33 4	1.27 4	1.18 4	1.16 4	1.70 4	0.93 4	1.52 4	1.55 5	0.81 3	3.27 5
Tower + QAD											
Tower-v2 70B+MBR	0.93 2	0.96 2	0.83 2	0.80 2	0.72 2	1.20 2	0.71 2	1.20 2	0.97 2	0.61 2	2.64 2
Tower-v2 70B+TRR	1.07 3	1.05 3	0.96 3	0.91 3	0.87 3	1.27 3	0.82 3	1.27 3	1.07 3	**0.59** 1	2.88 3
Tower-v2 70B 2-step	**0.91** 1	**0.94** 1	**0.77** 1	**0.76** 1	**0.70** 1	**1.14** 1	**0.68** 1	**1.17** 1	**0.94** 1	**0.57** 1	**2.59** 1

Table 1: Translation quality (via MetricX-QE-XXL) on the WMT24 test set. Tower-v2 with MBR/TRR ranks first across all language pairs. Even with Greedy decoding Tower-v2-70B still ranks above other strong systems like Claude-Sonnet-3.5, GPT-4o and DeepL except in en→hi and ja→zh where Claude-Sonnet-3.5 has similar scores.

formed full document translations using previous Tower models while controlling for translation quality using CometKiwi (Rei et al., 2022). At the end, we are left with more data for document-level than segment-level, further contributing to improved performance on paragraph- and document-level translation.

Model suite. Tower-v2 now comes in two sizes: a 7B parameter model based on Mistral-7B (Jiang et al., 2023) and a larger 70B model based on LLama-3-70B (AI@Meta, 2024).

4 Quality-aware decoding with Tower-v2

On LLM-based MT, translations are typically generated through lightweight decoding strategies such as greedy or nucleus sampling. Nevertheless, strategies informed by quality metrics such as Minimum Bayes Risk Decoding (MBR) and Tuned Reranking (TRR) consistently perform better compared to other methods (Fernandes et al., 2022; Freitag et al., 2022; Nowakowski et al., 2022; Farinhas et al., 2023). As such for our submission, we experiment with MBR and TRR. For both methods, we use a candidate pool of 100 samples and ϵ-sampling (Freitag et al., 2023a) with $\epsilon = 0.02$, and Comet22 as the target objective. For TRR, we use the WMT23 test set for tuning the weights[3]. The translation quality features used include: model log probabilities, Comet-QE-20, CometKiwi22, CometKiwi-XL, and xComet-QE-XL.

To leverage the strengths of both approaches, we also experiment with a second step of refinement. After obtaining translations from both MBR and TRR, we select the TRR translation only if all quality features (except the model log probabilities) agree that the TRR translation is better than the MBR translation; otherwise, we retain the MBR translation[4].

5 Experimental Setup

5.1 Evaluation Setup

During the development of Tower-v2, we used WMT23 as our validation set. For our final analysis, we use WMT24 test set source sentences and report only QE metrics: CometKiwi-XXL (Rei et al., 2023), MetricX-QE-XXL (Juraska et al., 2023), and xComet-QE-XXL (Guerreiro et al., 2023). Additionally, we add the official preliminary results to the Appendix which include MetricX (reference-based) (Kocmi et al., 2024b).

We use evaluation metrics to develop and op-

[3] We sample 5000 sentences from the WMT23 test set to train the weights more efficiently.

[4] According to both automatic and human evaluation (Table 2 and Table 3 respectively) results of MBR translations are generally better.

	en→xx			xx→yy		
Models	MetricX ↓	xComet↑	CometKiwi ↑	MetricX ↓	xComet↑	CometKiwi ↑
Baselines						
NLLB-54B	7.61 7	66.90 7	57.01 7	7.74 8	48.21 6	56.14 7
GPT-4o	1.50 6	83.74 6	77.04 5	2.18 5	70.44 2	76.19 4
Claude-Sonnet-3.5	1.40 5	84.85 5	78.09 4	1.98 4	69.73 2	76.77 4
DeepL	—	—	—	4.38 6	56.19 4	68.33 6
Tower						
Tower-v2 7B	1.48 5	83.77 5	77.02 5	2.24 5	67.44 4	75.86 4
Tower-v2 70B	1.32 4	84.87 4	78.29 4	2.04 4	69.20 3	76.70 4
Tower + QAD						
Tower-v2 70B+MBR	0.92 2	88.78 2	81.39 3	1.62 2	69.88 2	78.28 2
Tower-v2 70B+TRR	1.03 3	87.95 3	82.13 2	1.73 2	**71.95** 1	79.38 2
Tower-v2 70B 2-step	**0.89** 1	**89.25** 1	**82.54** 1	**1.58** 1	70.85 2	**79.69** 1

Table 2: Translation quality aggregated by language pairs on the WMT24 test set (without testsuites). We omit DeepL from the en→xx averages because it does not support two language pairs. All metrics are their XXL variant.

timize our models (e.g., using MBR and/or TRR during inference), with the exception of metrics of the MetricX family. Thus, to mitigate potential biases, we report MetricX-QE-XXL as our main evaluation metric and conduct human evaluation for English→German and English→Chinese. For the human evaluation, we use SQM quality levels with full document context. The annotators are in-house expert linguists familiar with evaluating MT outputs.

On Table 1, we report performance clusters based on statistically significant performance gaps at a 95% confidence threshold. On Table 2, we create per-language groups for systems with similar performance, following Freitag et al. (2023b), and obtain system-level rankings using a normalized Borda count (Colombo et al., 2022), which is defined as an average of the obtained clusters.

Regarding baselines, we report three commercial systems, GPT-4o, Claude-Sonnet-3.5, and DeepL, along with an open-source NMT model, NLLB 54B. While little is known about the commercial systems, they show top performance on the WMT23. All models are evaluated in a 0-shot setting, unless stated otherwise.

5.2 Main Results

Table 1 shows our main results on English→X language pairs according to MetricX-QE-XXL (↓). Table 2 shows aggregated scores for English→X and X→Y according to different metrics. From Table 1, we observe that even the 7B model

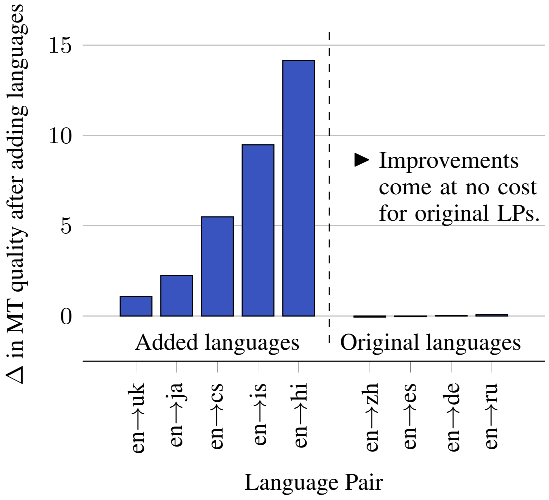

Figure 1: Improvement in MT quality after adding new languages to Tower-v2; measured in negative MetricX-XXL-QE so taller bars equate to better quality.

with greedy decoding outperforms, or is on par, with the best baseline, Claude-Sonnet-3.5, for English→X. Scaling to 70B brings consistent improvements across all language pairs, and both TRR and MBR decoding bring MetricX-QE-XXL further down. Our final submission (2-step) ranks first for all language pairs with statistical significance.

5.3 Impact of Adding 5 Languages

To evaluate the impact of adding 5 languages, we train two 7B models: one with the initial 10 languages of Tower; another with the 10 languages

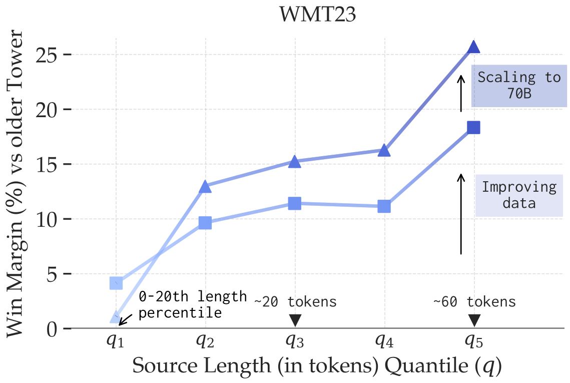

Figure 2: Win rates margin by length of the tokenized source of TOWER-V2-7B (squares) and TOWER-V2-70B (triangles) against an older iteration that was not trained on long-context translation training data. All language pairs of the WMT23 dataset that intersect with WMT24 are considered. We define a (sentence-level) win if the delta between two systems is superior to 1×10^{-3} METRICX-XXL points

plus Hindi, Japanese, Ukrainian, Czech, and Icelandic. The data distribution for CPT remains unchanged, but we increase the number of training tokens of the second model to accommodate the additional languages. For SFT, we extend the dataset by incorporating human-translated data from several sources.

Figure 1 illustrates the absolute difference in 0-shot translation quality between the two models. As expected, the model with additional support performs considerably better on the new languages[5]. Perhaps more interestingly, its performance on the initially supported languages — which is already state-of-the-art (Table 1) — remains largely unchanged.

5.4 Beyond sentence by sentence translation

Figure 2 compares the new versions of TOWER-V2 (7B and 70B) with an older TOWER version that had yet to be trained on data specifically tailored to improve long-context translation. Not only do TOWER-V2 models vastly outperform the older version, but the quality gap widens as source length increases.

Further to this point, we created a paragraph-level version of the WMT23 dataset, by joining segments of the same document into paragraphs with at most 4 sentences. Results in Table 4 show that our final models are considerably better at translating paragraphs than their older counterpart.

Decoding	en→de	en→zh
Batch 1		
Greedy	85.43	84.11
TRR	87.16	85.55*
MBR	88.50*	85.47*
Batch 2		
TRR	—	68.55
MBR	—	72.76*

Table 3: SQM quality evaluation for three different decoding methods using TOWER-V2 70B. Numbers marked with an asterisk (*) are statistically significant. For English→Chinese, since the results of the first batch were not significant, we conducted a second batch comparison between TRR and MBR.

5.5 Putting all together into 70B parameters

The gains from scaling up the number of parameters are clear from Tables 1 and 2, where we show that TOWER-V2-70B consistently outperforms all baselines in all language pairs, except ja→zh. Coupling TOWER-V2-70B with QAD methods yields state-of-the-art results for all languages and metrics considered. Remarkably, Figure 2 shows that the 70B model considerably improves upon its 7B counterpart suggesting that the benefits of scaling up are particularly noticeable when translating longer sources.

5.6 Human Evaluation: Greedy vs TRR vs MBR

To validate our findings with automatic metrics, we conducted a small-scale human evaluation for English→German and English→Chinese (Table 3). In a first phase, linguists annotated 100 samples from TOWER-V2-70B with different decoding strategies on the WMT24 test. For both language pairs, annotators scored greedy decoding lower than the other two methods. While there was a noticeable quality difference between MBR and TRR for English→German, this distinction was not evident for English→Chinese, with both decoding strategies achieving similar results. Therefore, we conducted a second round of annotations for English→Chinese, comparing only TRR with MBR. This provided more concrete results that favored MBR outputs.

[5] We note that the initial version of TOWER has ability to translate to other languages outside the supported ones, especially when given few-shot examples (Richburg and Carpuat, 2024) Still, their zero-shot performance is weak for languages like Hindi or Icelandic, which are less represented in the pretraining of the base models like LLaMA-2.

	WMT23-Paragraphs					
	en→xx			xx→yy		
Models	MetricX ↓	Comet ↑	chrF ↑	MetricX ↓	Comet ↑	chrF ↑
Tower (older)	5.14	79.11	50.93	6.99	75.45	53.29
Tower-v2-7B	2.72	84.45	54.35	1.87	87.57	61.36
Tower-v2-70B	**2.40**	**84.87**	**55.06**	**1.72**	**87.75**	**62.29**

Table 4: Performance of different Tower versions on our paragraph-level version of WMT23 (measured by MetricX-XXL, Comet-22, and chrF). Tower (older) is a version prior to the interventions we ultimately made on the training data of Tower-v2 to make it better at translating longer sources. These changes led to major improvements in paragraph-level translation for Tower-v2-7B, which are further realized with Tower-v2-70B.

5.7 Context-aware translation

	en→xx	
Models	MetricX ↓	xComet↑
Tower-v2-70B 0-shot	0.510	96.96
Tower-v2-70B 5-shot	0.495	96.89
	xx→en	
Tower-v2-70B 0-shot	1.051	94.84
Tower-v2-70B 5-shot	0.766	95.54

Table 5: Translation quality of Tower-v2-70B on the development set of the WMT24 Chat Shared Task. Using a prompt that incorporates conversational context (see Appendix A), the model provides high-quality translations, especially with examples (5-shot).

To evaluate Tower-v2 in a different domain, we tested it on chat translation data. In this domain, the model translates a segment based on the context of previous conversation turns. Ignoring this context can result in subpar translations with pronoun mistakes and lexical inconsistencies (Läubli et al., 2018; Toral et al., 2018). Table 5 shows that Tower-v2-70B excels at chat translation, even without specific training for this task. Using the prompt in Appendix A, which includes the conversation context, the model provides high-quality translations, especially when given domain-specific examples.

6 Conclusion

In this paper, we describe the joint submission from Unbabel and IST to the WMT24 General MT shared task. Our new model, Tower-v2, significantly improves upon previous versions by expanding language coverage from 10 to 15 languages and enhancing translation quality for longer paragraphs. Our largest model, with 70 billion parameters, combined with QAD strategies, achieved first place on the WMT24 test set according to both reference-free automatic evaluation, which we employed, and reference-based evaluation, as reported in the preliminary results from the WMT24 organizers (Kocmi et al., 2024b).

Limitations

This paper highlights the key improvements in Tower-v2 compared to previous versions and benchmarks it against other commercial state-of-the-art systems like GPT-4o, Claude-Sonnet-3.5, and DeepL. However, our submission is "unconstrained and closed," meaning the information provided is not sufficient for full system replication. Furthermore, our comparisons primarily focus on translation quality and do not consider factors like inference speed, training budget, or model efficiency.

We also disclose the number of parameters in our models, from the 7B version to the final 70B version, to facilitate a clearer understanding of their scale. However, these comparisons with other systems do not account for differences in model parameters and other operational metrics.

Acknowledgements

Part of this work was supported by the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

AI@Meta. 2024. Llama 3 model card.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre 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, Singapore. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022. What are the best systems? new perspectives on nlp benchmarking. In *Advances in Neural Information Processing Systems*.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10053–10071. PMLR.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023a. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023b. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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, 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. 2024a. Findings of the 2024 conference on machine translation (WMT24). In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida. 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, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024c. Error span annotation: A balanced approach for human evaluation of machine translation.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*.

Artur Nowakowski, Gabriela Pałka, Kamil Guttmann, and Mikołaj Pokrywka. 2022. Adam Mickiewicz University at WMT 2022: NER-assisted and quality-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 326–334, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Raphael Reinauer, Patrick Simianer, Kaden Uhlig, Johannes E. M. Mosig, and Joern Wuebker. 2023. Neural machine translation models can learn to be few-shot learners.

Aquia Richburg and Marine Carpuat. 2024. How multilingual are large language models fine-tuned for translation?

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Marcos Treviso, Nuno M. Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and André F. T. Martins. 2024. xtower: A multilingual llm for explaining and correcting translation errors.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Metrics for QAD

The translation quality features used include: model log probabilities, COMET-QE-20[6], COMETKIWI22[7], COMETKIWI-XL[8], and XCOMET-QE-XL[9].

A.2 Chat Translation Prompt

Given a source (SRC) to be translated from SRC_LANG to TGT_LANG, and previous turns

[6] Unbabel/wmt20-comet-qe-da

[7] Unbabel/wmt22-cometkiwi-da

[8] Unbabel/wmt23-cometkiwi-da-xl

[9] Unbabel/XCOMET-XL

in a conversation between two agents (TURN_i), the 0-shot prompt used was:

Context: <TURN_1>\n <TURN_2>. . . \n
<TURN_k>.\n\nTranslate the
<SRC_LANG>source text to <TGT_LANG>,
given the context.\n<SRC_LANG>:
<SRC>\n<TGT_LANG>:

When using five in-context examples, the prompt is repeated six times separated by two new lines; five times with a reference translation at the end, and one times exactly as written above.

A.3 Further analysis on long-context translation

Compared to the first version of Tower, the ability of Tower-v2 to translate long sources has greatly improved. Whereas the translation quality of latter fell behind GPT-4 for longer sources, Tower-v2-70B is superior across the board compared to the current best closed model for translation, Claude-Sonnet-3.5. In fact, the performance gap tends to widen as source length increases. Tower-v2-7B is also competitive for the first 4 quantiles of length, but falls slightly behind on the last one.

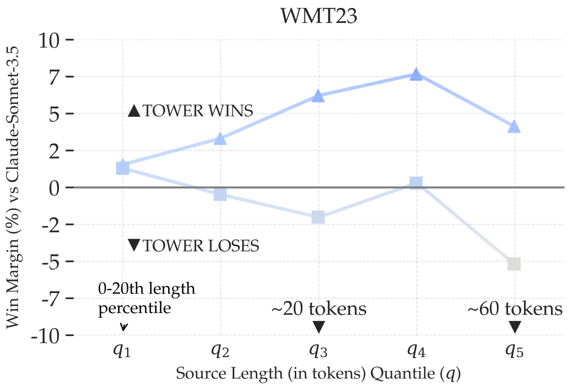

Figure 3: Win rates margin by length of the tokenized source of Tower-v2-7B (squares) and Tower-v2-70B (triangles) against Claude-Sonnet-3.5. All language pairs of the WMT23 dataset that intersect with WMT24 are considered. We define a (sentence-level) win if the delta between two systems is superior to 1×10^{-3} MetricX-XXL points

A.4 Preliminary Results from Kocmi et al. (2024b)

See Tables 6 to 16 for the official automatic evaluation conducted by WMT 24 organizers. Our submission, Unbabel-Tower70B, ranks first on all language pairs and metrics.

Czech-Ukrainian

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	0.9	0.719	✓
Claude-3.5 §	1.7	1.0	0.683	✓
IOL-Research	1.9	1.3	0.681	✓
CommandR-plus §	1.9	1.3	0.677	✓
GPT-4 §	2.0	1.4	0.677	✓
Gemini-1.5-Pro	2.0	1.2	0.668	✓
ONLINE-W	2.3	1.4	0.661	✓
Mistral-Large §	2.3	1.6	0.666	
IKUN	2.3	1.6	0.664	✓
Aya23	2.5	1.9	0.665	✓
TranssionMT	2.6	1.5	0.648	
ONLINE-B	2.6	1.6	0.648	
ONLINE-A	2.6	1.5	0.647	
Llama3-70B §	2.6	2.0	0.661	
ONLINE-G	2.8	1.8	0.639	
CUNI-Transformer	3.0	2.0	0.639	✓
IKUN-C	3.0	2.4	0.648	✓
Phi-3-Medium §	9.1	6.5	0.425	
BJFU-LPT †	11.5	7.6	0.321	
CycleL	21.0	19.5	0.146	

Table 6: Preliminary WMT24 General MT automatic ranking for Czech-Ukrainian.

English-Czech

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	1.8	0.732	✓
Claude-3.5 §	2.1	2.4	0.693	✓
CUNI-MH	2.1	2.3	0.690	✓
CUNI-GA	2.3	3.7	0.726	✓
Gemini-1.5-Pro	2.6	2.8	0.678	✓
GPT-4 §	2.6	2.9	0.682	✓
IOL-Research	2.8	3.0	0.676	✓
ONLINE-W	2.8	2.8	0.669	✓
CommandR-plus §	2.9	2.9	0.669	✓
SCIR-MT	3.2	3.3	0.664	✓
TranssionMT	3.5	3.5	0.655	
ONLINE-A	3.6	3.4	0.648	
Mistral-Large §	3.7	3.6	0.647	
IKUN	3.9	3.7	0.638	✓
ONLINE-B	4.0	3.9	0.640	
Llama3-70B §	4.1	4.0	0.640	✓
Aya23	4.3	4.0	0.630	✓
CUNI-DocTransformer	4.4	4.0	0.621	✓
IKUN-C	4.7	4.3	0.618	✓
CUNI-Transformer †	4.7	4.3	0.614	
ONLINE-G	5.7	5.2	0.592	
NVIDIA-NeMo †	7.6	6.5	0.536	
Phi-3-Medium §	15.0	11.4	0.305	
TSU-HITs	19.5	16.6	0.235	
CycleL2	24.2	19.5	0.077	
CycleL	27.0	22.5	0.031	

Table 7: Preliminary WMT24 General MT automatic ranking for English-Czech.

English-German

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	1.1	0.723	✓
Dubformer	1.8	1.2	0.694	✓
TranssionMT	1.8	1.4	0.699	✓
GPT-4	1.8	1.4	0.700	✓
ONLINE-B	1.8	1.4	0.698	✓
Claude-3.5	1.9	1.4	0.695	✓
CommandR-plus	2.0	1.4	0.696	✓
Mistral-Large	2.0	1.5	0.694	✓
Gemini-1.5-Pro	2.2	1.5	0.688	✓
ONLINE-W	2.2	1.5	0.689	
IOL-Research	2.3	1.6	0.692	✓
Llama3-70B §	2.5	1.7	0.686	✓
Aya23	2.7	1.8	0.680	✓
IKUN	3.0	1.8	0.668	✓
ONLINE-A	3.0	1.8	0.667	
Phi-3-Medium §	3.4	2.0	0.657	
ONLINE-G	3.5	2.1	0.662	
IKUN-C	3.8	2.0	0.641	✓
CUNI-NL	4.2	2.1	0.624	
AIST-AIRC	7.2	3.3	0.551	
NVIDIA-NeMo †	7.4	3.5	0.558	
Occiglot	8.2	3.8	0.539	
MSLC	11.9	4.4	0.390	
TSU-HITs	13.3	5.6	0.395	
CycleL2	27.0	11.5	0.091	
CycleL	27.0	11.5	0.091	

Table 8: Preliminary WMT24 General MT automatic ranking for English-German.

English-Spanish

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	1.9	0.745	✓
GPT-4	1.9	2.5	0.712	✓
Dubformer	2.0	2.2	0.700	✓
CommandR-plus	2.1	2.6	0.706	✓
Claude-3.5	2.1	2.6	0.705	✓
Mistral-Large	2.2	2.7	0.707	✓
IOL-Research	2.3	2.8	0.701	✓
Gemini-1.5-Pro	2.4	2.8	0.696	✓
Llama3-70B §	2.6	3.0	0.693	✓
ONLINE-B	2.7	3.1	0.690	
ONLINE-W	2.7	3.0	0.682	
TranssionMT	2.8	3.2	0.689	
IKUN	2.8	3.3	0.687	✓
Phi-3-Medium §	3.0	3.4	0.685	
ONLINE-A	3.0	3.3	0.676	
Aya23	3.1	3.5	0.681	
ONLINE-G	3.2	3.6	0.674	
IKUN-C	3.4	3.5	0.666	✓
NVIDIA-NeMo †	4.5	4.4	0.631	
Occiglot	5.9	5.4	0.583	
MSLC	7.4	6.4	0.532	✓
TSU-HITs	16.3	14.2	0.289	
CycleL	24.0	20.9	0.072	

Table 9: Preliminary WMT24 General MT automatic ranking for English-Spanish.

English-Hindi

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	3.1	0.657	✓
Claude-3.5 §	1.2	3.3	0.649	✓
TranssionMT	1.3	3.3	0.644	✓
ONLINE-B	1.4	3.3	0.641	✓
Gemini-1.5-Pro §	1.6	3.6	0.635	✓
GPT-4 §	2.1	4.5	0.628	✓
IOL-Research	2.1	4.3	0.622	✓
Llama3-70B §	2.1	4.6	0.630	✓
CommandR-plus §	2.3	4.4	0.612	
Aya23	3.2	5.4	0.591	✓
ONLINE-A	3.5	6.2	0.590	
ONLINE-G	4.2	7.4	0.583	
Mistral-Large §	5.0	7.7	0.541	
IKUN-C	5.5	7.1	0.499	✓
NVIDIA-NeMo †	5.8	8.9	0.530	
Phi-3-Medium §	7.4	10.7	0.483	
IKUN	7.7	9.4	0.428	
ONLINE-W	15.3	20.9	0.296	
CycleL	20.0	23.4	0.083	

Table 10: Preliminary WMT24 General MT automatic ranking for English-Hindi.

English-Icelandic

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	2.5	0.740	✓
Claude-3.5 §	2.3	3.6	0.697	✓
Dubformer	2.5	3.4	0.685	✓
IKUN	3.2	4.3	0.666	✓
GPT-4	3.4	4.7	0.673	✓
AMI	3.7	4.9	0.663	✓
IKUN-C	3.7	4.9	0.657	✓
TranssionMT	4.2	5.5	0.653	
ONLINE-B	4.2	5.5	0.652	
IOL-Research	4.3	5.7	0.655	✓
ONLINE-A	5.5	6.4	0.603	
Llama3-70B §	6.7	8.0	0.586	✓
ONLINE-G	6.9	7.9	0.573	
CommandR-plus §	9.8	10.6	0.487	
Mistral-Large §	10.4	10.9	0.465	
Aya23 §	15.2	14.9	0.311	
Phi-3-Medium §	16.2	15.7	0.278	
ONLINE-W	18.1	19.5	0.296	
TSU-HITs	19.2	18.4	0.192	
CycleL	21.0	20.2	0.148	

Table 11: Preliminary WMT24 General MT automatic ranking for English-Icelandic.

English-Japanese

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	2.0	0.762	✓
ONLINE-B	1.4	2.4	0.750	✓
Claude-3.5	1.5	2.3	0.744	✓
Gemini-1.5-Pro	1.7	2.5	0.734	✓
GPT-4	1.7	2.7	0.740	✓
Team-J	1.9	2.9	0.740	✓
NTTSU	1.9	2.6	0.731	✓
CommandR-plus	1.9	2.7	0.730	✓
IOL-Research	2.3	3.1	0.724	✓
Aya23	2.3	3.1	0.719	✓
Llama3-70B §	2.6	3.5	0.714	✓
DLUT-GTCOM	2.6	3.0	0.697	
Phi-3-Medium §	2.8	3.6	0.709	
ONLINE-W	2.9	3.6	0.700	
Mistral-Large §	2.9	3.8	0.707	
ONLINE-A	3.0	3.6	0.699	
IKUN	3.1	3.7	0.696	
IKUN-C	3.9	4.3	0.669	✓
ONLINE-G	6.4	6.6	0.599	
AIST-AIRC	6.6	6.5	0.583	
UvA-MT	6.7	6.7	0.589	
NVIDIA-NeMo †	6.9	6.9	0.582	
CycleL	24.0	22.4	0.101	

Table 12: Preliminary WMT24 General MT automatic ranking for English-Japanese.

English-Russian

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	2.4	0.742	✓
Dubformer	1.9	2.8	0.701	✓
Yandex	1.9	2.9	0.705	✓
Claude-3.5	2.0	3.0	0.706	✓
ONLINE-G	2.2	3.3	0.706	✓
GPT-4	2.3	3.4	0.703	✓
Gemini-1.5-Pro	2.3	3.2	0.697	✓
CommandR-plus §	2.4	3.4	0.693	✓
ONLINE-W	2.6	3.5	0.688	
IOL-Research	2.6	3.7	0.694	✓
Mistral-Large §	2.7	3.7	0.692	
Llama3-70B §	3.1	4.1	0.681	✓
ONLINE-B	3.1	3.9	0.673	
TranssionMT	3.1	3.9	0.673	
IKUN	3.2	4.1	0.675	✓
Aya23	3.3	4.2	0.669	✓
ONLINE-A	3.4	4.1	0.663	
Phi-3-Medium §	3.9	4.7	0.654	
IKUN-C	3.9	4.7	0.649	✓
CUNI-DS	5.9	6.2	0.584	
NVIDIA-NeMo †	7.2	7.3	0.549	
TSU-HITs	10.8	9.8	0.421	
CycleL	24.3	22.2	0.062	
CycleL2	25.0	22.4	0.027	

Table 13: Preliminary WMT24 General MT automatic ranking for English-Russian.

English-Ukrainian

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	2.2	0.732	✓
Dubformer	1.8	2.7	0.691	✓
Claude-3.5	2.0	3.0	0.693	✓
ONLINE-W	2.1	2.8	0.679	✓
Gemini-1.5-Pro	2.2	3.0	0.677	✓
CommandR-plus §	2.3	3.2	0.678	✓
GPT-4	2.3	3.3	0.682	✓
ONLINE-G	2.3	3.1	0.670	
IOL-Research	2.4	3.4	0.675	✓
Mistral-Large §	2.4	3.4	0.675	
IKUN	2.8	3.7	0.661	✓
ONLINE-B	3.1	3.9	0.646	
TranssionMT	3.1	4.0	0.646	
Llama3-70B §	3.2	4.2	0.647	
Aya23	3.3	4.2	0.642	
ONLINE-A	3.3	4.1	0.634	
IKUN-C	3.9	4.7	0.622	✓
NVIDIA-NeMo †	6.2	7.0	0.537	
Phi-3-Medium §	11.1	11.3	0.339	
CycleL	21.0	22.4	0.037	

Table 14: Preliminary WMT24 General MT automatic ranking for English-Ukrainian.

English-Chinese

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	2.3	0.726	✓
Claude-3.5	1.7	3.0	0.703	✓
ONLINE-B	1.7	2.9	0.697	✓
IOL-Research	1.8	3.1	0.700	✓
Gemini-1.5-Pro	1.8	3.1	0.698	✓
GPT-4	2.0	3.3	0.693	✓
CommandR-plus	2.2	3.3	0.681	✓
ONLINE-W	2.2	3.2	0.677	
HW-TSC	2.3	3.4	0.675	✓
Mistral-Large §	2.8	4.0	0.665	
Llama3-70B §	2.8	3.9	0.662	✓
Aya23	3.0	4.1	0.655	✓
IKUN	3.1	4.0	0.646	✓
Phi-3-Medium §	3.1	4.2	0.648	
ONLINE-A	3.3	4.1	0.636	
IKUN-C	3.5	4.2	0.624	✓
UvA-MT	4.3	5.2	0.607	
ONLINE-G	4.8	5.5	0.588	
NVIDIA-NeMo †	7.3	7.6	0.494	
CycleL	20.1	20.1	0.086	
CycleL2	22.0	22.1	0.030	

Table 15: Preliminary WMT24 General MT automatic ranking for English-Chinese.

Japanese-Chinese

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	3.2	0.622	✓
Claude-3.5	1.7	3.5	0.603	✓
Gemini-1.5-Pro	1.9	3.5	0.595	✓
DLUT-GTCOM	2.0	3.3	0.586	✓
GPT-4	2.1	3.8	0.597	✓
IOL-Research	2.2	3.9	0.593	✓
CommandR-plus	2.8	4.1	0.576	✓
Team-J	2.8	4.0	0.570	✓
Llama3-70B §	3.1	4.7	0.578	✓
Mistral-Large §	3.5	4.9	0.568	
Aya23	3.7	5.0	0.563	✓
NTTSU	3.7	5.3	0.566	✓
Phi-3-Medium §	4.0	5.1	0.552	
IKUN	4.4	5.4	0.544	✓
ONLINE-B	5.2	5.5	0.518	
UvA-MT	5.2	6.3	0.534	
ONLINE-W	5.3	6.0	0.522	
IKUN-C	5.5	6.2	0.519	✓
ONLINE-A	6.8	6.8	0.484	
MSLC	8.9	8.8	0.450	
ONLINE-G	10.3	9.6	0.413	
CycleL	23.0	21.5	0.202	

Table 16: Preliminary WMT24 General MT automatic ranking for Japanese-Chinese.