

Relay-MT: Multi-agent LLM Ensembling for Machine Translation – Samsung R&D Poland Submission for WMT2026

Adam Dobrowolski¹, Paweł Przybysz¹, Paweł Przewłocki¹,
Marcin Szymański¹, Artur Janicki²

¹ Samsung R&D Institute Poland, ² Warsaw University of Technology,
{a.dobrowols2@samsung.com, p.przybysz@samsung.com
p.przewlocki@partner.samsung.com, m.szymanski@samsung.com
artur.janicki@pw.edu.pl}

Abstract

We propose a method to improve machine translation (MT) quality by ensembling multiple large language models (LLMs) within a multi-agent framework. The method addresses the problem of ensembling models with different tokenizations by abstracting away from token-level representations and operating at the level of generated text. The approach integrates several established concepts in MT, including ensembling, quality estimation, and reranking, into a single translation system. The log-likelihood reranking applied in this paper significantly improves results, even without ensembling. Experimental results show that the total improvements of the method are comparable to those of COMET-MBR.

1 Introduction

Previous approaches to WMT workshops have focused mainly on training the best possible models. Teams at leading laboratories are building increasingly effective machine translation (MT) solutions, and it is difficult to develop something superior. In the approach presented here, we took a different strategy. Rather than building a competitive model, we decided to leverage and combine the power of the best models currently available to the public.

We propose the Relay-MT algorithm described in Section 3 that allows for the ensembling of LLMs from different research teams, regardless of their tokenization and architecture. The presented method differs from other approaches in that it does not attempt to reconcile and align tokens across different models, which is a challenging task. Instead, we exchange information between models by passing generated hypotheses purely at the text level.

We compared the results of the presented algorithm against standard Minimum Bayes Risk decoding (Kumar and Byrne, 2004) with chrF (Popović, 2015) and COMET (Rei et al., 2020) as metrics.

2 Related work

In this section, we briefly describe related work on selected techniques frequently used in MT, which, when combined, form the algorithm presented in this paper.

2.1 Ensembling

Ensembling is a common technique that combines predictions from multiple models. In MT, ensembles generate text using several models in parallel. During decoding, each model in the ensemble calculates its own probabilities for the next tokens, which are then combined using experimentally chosen weights. At each step, the token with the highest weighted average probability across the models is selected. By coordinating multiple models, low-probability tokens that might be mistakenly selected due to a single model’s error are filtered out.

The rapid development and availability of high-quality LLMs have recently led to significant improvements in translation quality, making the application of ensemble methods a natural next step. However, traditional ensemble methods in MT face important limitations in the era of LLMs. The most significant limitation is the varying tokenization of LLMs. This means models operate on different tokens, making it challenging to compute average probabilities when tokens from different models appear in different positions within the text.

Several recent studies have explored various approaches to tackle this problem. For example, Kasai et al. (2022) and Wicks et al. (2025) address the issue of different tokenizations by proposing methods for mapping tokens between models through the analysis of dependencies between their text representations. The models align tokens with similar textual representations whenever possible, and probabilities are then recalculated to prepare better translations based on agreed common predictions.

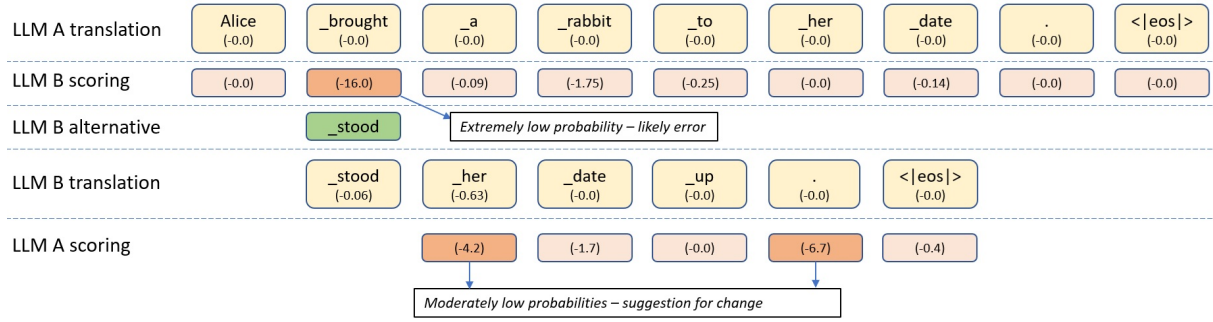


Figure 1: Translation process of French « Alice a posé un lapin à son rendez-vous. » (idiomatic “Alice didn’t show up for her date.”) to English with two LLMs. Detection of possible errors → translation from error position with another LLM. Values in parentheses are logprobs of tokens.

Xu et al. (2024) propose a mathematically advanced method that embeds tokens from different vocabularies into one shared space. Tokens are then combined and compared within this common representation.

These approaches still use a classic token-by-token hypothesis generation approach, offering methods for unifying tokens across multiple models.

2.2 Reranking

Reranking is a method that improves translation quality by using models other than the generating model to evaluate it. The role of one model is to generate multiple translation hypotheses, which are evaluated by other models. Thanks to verification across multiple models, we can select a better translation than a single model would produce. This method is used, for example, by the xTower model (Treviso et al., 2024) or the Chimera version of Hunyuan-MT (Zheng et al., 2025b).

The concepts presented here are inspired by Junczys-Dowmunt (2019), who presented a method that uses cross-entropy of sentence pairs to effectively distinguish correct translations from incorrect ones. This approach enables the detection of translation errors using token-level probability values produced by translation models. The results of that work also suggest that high sentence-level cross-entropy is driven by high cross-entropy at the individual token level, a hypothesis that served as one of the foundations for the error detection mechanism in the Relay-MT algorithm proposed in this paper.

While analyzing generated translations, we observed that low token probabilities consistently correlate with translation errors. This is expected behavior. It is unlikely that two different models

would assign high probability to the same incorrect token.

A necessary part of the reranking is developing a method for quickly generating a large number of hypotheses for subsequent evaluation. The primary method for generating multiple hypotheses is *beam search*, in which we simultaneously generate several highly probable translations. However, this method generates very similar hypotheses and overlooks many potentially better candidates. To increase the diversity of beam search, the *Diverse Beam Search* modification was developed (Vijayakumar et al., 2018).

Another popular method for generating multiple hypotheses is *sampling*, where, during text generation, we do not select the most probable token but instead allow for random selection from among the less probable ones. This method is simple and effective, but sometimes produces incorrect results. The best of sampled hypotheses can then be selected using Minimum Bayes-Risk Decoding (Kumar and Byrne, 2004).

Recent work has also focused on generating more diverse and higher-quality hypotheses. Vernikos and Popescu-Belis (2024) proposed cutting the set of generated hypotheses into parts (*QE-Fusion*) and stitching well-translated fragments together into new sentences. This method divides the translation into spans, as in Relay-MT, but operates on hypotheses from a single model.

Dobrowolski et al. (2025) proposed *A* decoding*, a method for fast generation of many diverse hypotheses by replacing most probable tokens with other, relatively likely ones, which yields a richer set of hypotheses than beam search or sampling.

Algorithm 1 Iterative multi-agent ensemble of models with different tokenizers

```
1: Initialize the base translation  $y$ 
2: Select the initial leading model  $M$ 
3: repeat
4:   Generate new hypothesis  $y'$  using model  $M$  from selected position in  $y$ 
5:   Compute token probabilities in  $y'$  using other models
6:   Find a new start position  $t = \arg \min_t p(y'_t)$ 
7:   Select another leading model  $M$  with the lowest score for token  $y'_t$ 
8:   Update the base translation  $y \leftarrow y'$ 
9: until no improvement in quality (evaluation function does not increase)
10: return  $y$ 
```

3 Relay-MT algorithm

Unlike regular approaches to ensembling in translation, this paper introduces a *multi-agent approach*. Here, model cooperation does not involve reconciling tokens generated at each step, but rather transferring the role of the primary generator among different models during translation. Generated hypotheses are passed between models as texts. With this agent-based approach, each model operates on its own tokens. This way, we avoid the problem of differing tokenization, which was a major issue when integrating models through classical ensembling.

Each LLM in the ensemble generates its own translation. The LLMs in the ensemble evaluate each other’s translations by checking the probabilities of tokens of the generated text. A low token probability indicates a likely error. In the place of the least probable token, the evaluating LLM takes over the role of the hypothesis generator, and the previous generator becomes the evaluator. In this way, a new translation is created, with its prefix generated by one LLM and its suffix by another. For this new translation, we repeat the process of searching for low-probability tokens and switching the role of the translation model. After several such role switches (relays), we obtain a new hypothesis created by combining several fragments generated by different LLMs and free of most likely errors.

By applying the algorithm multiple times, we can obtain any number of alternative, high-quality hypotheses. These hypotheses are more diverse than those obtained from sampling a single model; at the same time, all text fragments are generated not randomly, like in sampling, but by a set of translation LLMs, ensuring high quality. The full version of the proposed generation pseudocode is shown in Algorithm 1.

4 Log-likelihood Reranking

Changes to the translation during the process correct some errors but may introduce new ones. To minimize the likelihood of errors in the final translation, we rerank the generated hypotheses at the end of the algorithm, selecting the best of all generated hypotheses.

The choice of the reranking method is independent of the Relay-MT method itself. Any reranking method can be used. The method described here is based on the conditional probability of a generated translation; therefore, it fits perfectly to Relay-MT as all necessary probabilities have already been calculated during mutual evaluation of translations between ensemble models and does not require any additional LLM computations.

All generated hypotheses are reranked within the ensemble by calculating the weighted average normalized log-probability of the generated sentences.

$$Q_{ens}(h|s) = \sum_i \lambda_i \frac{1}{|h|} \log P_i(h|s) \quad (1)$$

where $P_i(h|s)$ represents the conditional probability assigned to hypothesis h by the i -th model of the ensemble, and λ_i controls its contribution to the final score. The highest-scoring hypothesis is selected as the final translation.

To select the optimal λ_i coefficients we applied a simple grid-search to select a set of coefficients that would provide the best translations of the blindset according to Q_{wmt} function described in Section 5.2 although any specific quality assessment (QE) method can be applied here.

5 WMT26 submission

In this section, we present the implementation details used to prepare translations for WMT26.

5.1 Single Step of Relay-MT

At the start of the algorithm, we translate the entire blindset using all models in the ensemble. This way, we produce basic, default translations of the blindset by all LLMs of the ensemble. Then, for each hypothesis, we run scorings using other models in the ensemble to identify the least probable tokens and the alternatives proposed by the evaluating models at these locations. The position where the difference between the alternative and original token is greatest is a potential place where we can improve the translation the most.

For each source text and each evaluating model, we select the hypothesis with the greatest difference between the alternative and original token probabilities. We prepare the prefix of this translation up to the position of a detected low-probability token, replacing it with its alternative. In the next step of the algorithm, we perform translation of the source text using the evaluating model, starting from the replacement position using the constructed prefix. Prefixes already translated are omitted.

The above steps are repeated several times to generate a set of hypotheses, which are finally reranked to select the best hypothesis. One iteration of the algorithm takes one step of translation and several steps of scoring for each model of the ensemble.

In its current form, Relay-MT is relatively computationally expensive. 10 iterations of the algorithm for 5,000 paragraphs involving three ~30B models take approximately 2,000 seconds on two NVIDIA RTX PRO 6000 Blackwell Max-Q GPUs. Comparable in quality (see Table 1) generation of 20 hypotheses by sampling, followed by COMET-MBR reranking, takes approximately 800 seconds, making Relay-MT roughly 2.5 times slower. On three NVIDIA A100-PCI-40GB GPUs, the corresponding execution times are approximately 5,000 and 2,400 seconds, respectively, making Relay-MT about twice slower.

5.2 WMT2024/25 tuned QE

The quality of translations in recent years has reached such a high level that even the Best-QE metrics, such as COMETKiwi, xCOMET, and MetricX, are no longer sufficient to compare hypotheses at this level. We have therefore introduced a more sophisticated QE function based on a similar linear combination of log-probabilities as Equation 1.

As P_i functions of the equation, we used log-probabilities calculated by various models: forward and backward translations, monolingual models and QE models. To make the function as close as possible to human ratings, we adjusted λ_i coefficients on the basis of the human evaluations from the results of WMT24 (Kocmi et al., 2024) and WMT25 (Kocmi et al., 2025b). We treated the human evaluations as ground truth and adjusted the weights so that the best-ranked hypotheses would give maximum average human rating. Finally, we selected the most relevant models, rounded the weights, and derived the following function:

$$Q_{wmt}(h|s) = 2P_{HY15}(h|s) + 2P_{HY30}(h|s) + P_{Qwen}(h|s) + 0.25P_{Tower}(h|s) + 0.25P_{Euro}(h|s) - 0.1P_{MetricX}(h|s) \quad (2)$$

The formula above contains normalized log-probabilities of forward translations by the following models:

- P_{HY15} - HY-MT1.5-7B
- P_{HY30} - Hy-MT2-30B-A3E
- P_{Qwen} - Qwen3.5-122B-A10E
- P_{Tower} - Tower-Plus-9B¹
- P_{Euro} - EuroLLM-22B¹
- $P_{MetricX}$ - value of MetricX-24

5.3 Document Segmentation

The source documents were divided into paragraphs, and each paragraph was translated separately. This division may have resulted in a loss of coherence in longer texts. However, translating very long texts, over 1000 tokens, was extremely time-consuming. Additionally, due to its nature, the Relay-MT algorithm makes only a few changes (equal to the number of iterations) to a given text, which would not significantly affect the long text output. After all paragraphs of a single document had been translated, the translations were recombined back into a single output document.

5.4 Model Selection

To build ensembles, we used the following translation models:

- Marco-MT-Algharb (Wang et al., 2025)
- EuroLLM-22B-Instruct (Martins et al., 2024)
- command-a-translate (Kocmi et al., 2025a)
- Hunyuan-MT-7B (Zheng et al., 2025b)
- HY-MT1.5-7B (Zheng et al., 2025a)

¹Only for directions supported by the model

- Hy-MT2-30B-A3B (Zheng et al., 2026)
- Hy-MT2-7B (Zheng et al., 2026)
- MiLMMT-46-12B (Shang et al., 2026)
- Qwen3.5-27B (Yang et al., 2025)
- Qwen3.5-122B-A10B (Yang et al., 2025)
- Qwen3.6-27B (Yang et al., 2025)
- Seed-X-PPO-7B (Cheng et al., 2025)
- Tower-Plus-9B (Rei et al., 2025)
- translategemma-27b (Finkelstein et al., 2026)

To find the best ensembles, we ranked the above models based on translation quality for a selected direction. We built ensembles consisting of 2–4 best models each and ran the Relay-MT algorithm for 5 steps to determine which ensembles produce the best results. We used the Q_{wmt} score to compare the quality between ensembles and selected the 4 best ensembles for each direction.

We ran 10 steps of the Relay-MT algorithm with these best ensembles, obtaining several dozens (Exactly: 10 times the number of models) hypotheses for each source paragraph. All of these hypotheses were reranked within an ensemble using Q_{ens} function, and the best hypothesis obtained in this way became the final translation proposal for the ensemble.

5.5 Final Error Correction

Despite the reranking function, some hypotheses contain errors, despite verification by the reranking function Q_{ens} . For example, translations that consist of many repetitions or are truncated in the middle elude the Q_{ens} verification, which yield good scores despite being incorrect. To eliminate such errors, we selected the hypotheses generated by the two best ensembles (not one-by-one) and compared them sentence-by-sentence using raw MetricX-24. The sentence from each pair that scored higher was selected as the final hypothesis and reported in the final result submitted to WMT26.

6 Results Achieved

A summary of the results for the proposed algorithm is presented in Table 1. The table contains average QE scores for the WMT26 blindset, calculated at the paragraph level and averaged across 13 main translation directions. Rows present evaluations of raw outputs of the 6 models used in ensembles, followed by the results of our experiments:

- *Rows labeled with (MBR)* – show results of MBR applied to 20 hypotheses sampled by a

Model	COMETKiwi	MetricX	xCOMET	ReMedy
HY-MT1.5-7B	66.3	-3.74	75.1	55.2
MiLMMT-46-12B	64.0	-4.55	71.2	49.4
Qwen3.5-27B	64.0	-4.43	68.7	48.8
Seed-X-PPO-7B	63.2	-4.70	69.3	43.8
command-a-translate	64.1	-4.71	69.9	44.4
translategemma-27b-it	67.0	-3.68	73.6	52.7
HY-MT1.5-7B (COMET-MBR)	66.9	-3.58	76.2	61.3
translategemma-27b (COMET-MBR)	67.6	-3.51	74.9	61.4
MultiLLM (chrF-MBR)	67.1	-3.65	76.0	62.9
MultiLLM (COMET-MBR)	68.3	-3.36	76.9	65.3
Log-likelihood reranking	67.1	-3.60	76.4	59.9
Best Relay-MT ensemble	67.3	-3.56	76.9	60.1
Our Submission	67.4	-3.47	77.1	62.5
Best-QE Ensemble	67.7	-3.28	77.3	65.7

Table 1: Summary of results of translations by single LLMs and MBR compared to the presented solution. Bold indicates the best result.

model with temperature=0.7

- *MultiLLM-MBR* – show results of MBR applied to 3x20 hypotheses sampled by three models: HY-MT1.5-7B, Hy-MT2-30B-A3B, translategemma-27b-it
- *Log-likelihood reranking* – shows QE values just after the first step of the Relay-MT algorithm, where basic translations are generated by all LLMs in an ensemble and reranked with the Q_{ens} function, before any error detection or relay of the translator role.
- *Best Relay-MT ensemble* – shows QE values for the best ensemble after 10 steps of the Relay-MT algorithm.
- *Our submission* – shows QE values after the *Final Error Correction* stage described in the Section 5.5.
- *Best-QE Ensemble* – shows results of Relay-MT for an ensemble with highest QE values.

As an official submission for WMT26 we provided the outcome of the ensemble not with the highest QE values but the ensemble best evaluated by Equation 2 which shows the best correlation to human evaluations (see section 5.2). Results of the best ensemble by QE presented in the last row of Table 2 shows that potential of Relay-MT in quality gains is comparable to COMET-MBR.

In Appendix A, we present detailed results for 13 translation directions separately.

7 Comparison to other participants

After receiving the results of all WMT26 participants from the organizers (Kocmi et al., 2026), we

System	COMET	MetricX ref	xCOMET ref	BLEU	chrF	COMET Kiwi	MetricX QE	xCOMET QE
Cohere_CAT+	0.8560	-6.9803	0.4935	27.2632	52.5191	0.5634	-6.8362	0.4221
UvA-MT	0.8510	-7.0541	0.4923	29.8434	55.0302	0.5685	-7.0185	0.4204
GPT_5.5	0.8508	-7.0829	0.5022	33.0217	56.9261	0.5696	-7.0971	0.4277
Wayfinder	0.8493	-7.0965	0.4964	29.6887	54.6077	0.5694	-6.9987	0.4258
TRIVE	0.8486	-7.2550	0.4932	30.7461	55.6318	0.5715	-7.0988	0.4265
Lumen	0.8484	-7.2001	0.4911	29.4306	54.1936	0.5673	-7.0793	0.4229
SRPOL (Relay-MT)	0.8477	-6.6062	0.5052	23.2542	49.0736	0.5920	-6.1205	0.4754
PseudoRef	0.8467	-6.6534	0.4912	27.5946	51.9028	0.5812	-6.2934	0.4458
translategemma-27b-it	0.8464	-7.0806	0.4759	23.1837	48.8621	0.5489	-7.1448	0.4121
HY-MT2-30B-A3B	0.8457	-7.0036	0.4930	22.9874	48.8809	0.5577	-7.0567	0.4357
Dubformer	0.8452	-7.6040	0.4827	29.1753	53.4750	0.5459	-7.5608	0.3986
MakotoAI	0.8410	-7.3645	0.4868	31.6025	55.3568	0.5673	-7.2347	0.4226
Gemini_3.1_Pro	0.8395	-7.6197	0.4847	29.6024	54.5405	0.5667	-7.3854	0.4178
Qingqiu-MT-9B	0.8382	-7.9929	0.4673	28.1590	52.9669	0.5650	-7.4075	0.4185
Command_A+_with-reasoning	0.8370	-7.7773	0.4663	24.8461	49.4592	0.5731	-7.0368	0.4353
GoogleTranslate	0.8343	-7.5129	0.4692	26.3086	50.4684	0.5636	-7.0424	0.4284
DeepSeek_V4_Pro_with-reasoning	0.8337	-7.7927	0.4742	29.1931	53.2355	0.5636	-7.3917	0.4218
MiLMMT-46-12B-v0.1	0.8306	-7.6702	0.4622	25.8864	49.7478	0.5408	-7.5615	0.4144
Qwen3.5-122B-A10B	0.8301	-7.8322	0.4600	24.1266	49.3609	0.5386	-7.7500	0.4032
Gemma_4_-_31B	0.8289	-8.5620	0.4566	26.6612	51.9729	0.5635	-7.6465	0.4153
GPT_OSS_120B	0.8177	-8.6426	0.4315	22.6254	47.0239	0.5501	-7.7614	0.4068
Mistral_Medium_3.5	0.8132	-8.7172	0.4346	24.4158	48.2901	0.5500	-7.8714	0.4014

Table 2: Results of the top evaluated multilingual systems submitted to WMT26 and raw outputs of LLMs. The SRPOL row reports the results of our Relay-MT solution. Cell colors represent relative performance within each metric column, ranging from light red (lowest value) to light green (highest value).

were able to perform an additional analysis and compare our system with those submitted by other teams.

The comparison is presented in Table 2. The columns report automatic scores according to different evaluation metrics, averaged over several translation directions (the same directions as reported in Appendix A). The table includes also results of translations by 4 LLMs not provided by organizers (translategemma-27b-it, HY-MT2-30B-A3B, Qwen3.5-122B-A10B, MiLMMT-46-12B-v0.1). These 4 models were base models of most our ensembles.

The SRPOL system (Relay-MT described in this paper) achieved the best results in 5 out of the 8 automatic evaluation metrics (MetricX, xCOMET, COMETKiwi, MetricX-QE, and xCOMET-QE). However, its BLEU/chrF scores are among lowest compared to those of other teams.

Our preliminary analysis suggests that these low scores are likely a consequence of the low BLEU/chrF performance of the base models on which we built the ensembles. These models achieve high QE scores, which are further improved by Relay-MT. However, the algorithm cannot compensate for the low baseline BLEU/chrF scores. Such low BLEU/chrF scores may negatively impact the final human evaluation. The rea-

son for this discrepancy is not yet known and requires further investigation.

8 Conclusion and Future Work

In this paper, we described the practical application of the Relay-MT algorithm for ensembling of translation models of different tokenization. The newly developed Relay-MT algorithm has demonstrated its effectiveness in producing high-quality translations. Results after 10 steps of the Relay-MT algorithm are comparable or better than the application of the standard Minimum Bayes Risk with COMET for 20 sampled hypotheses. The final results show exceptionally large gains according to the QE metrics, in stark contrast to low BLEU/chrF scores. This discrepancy requires further investigation. An unexpected finding of our research is the strong performance of the log-likelihood reranking method described in Section 4, which significantly improves translation results after only a single translation without the Relay-MT algorithm.

The proposed method provides full control over the generated translations, making it suitable for practical applications. Concepts of the Relay-MT algorithm, such as ensembling multiple models, log-likelihood reranking, and error detection based on low token probabilities, could serve as a basis for further research to improve translation quality.

Limitations

We presented only the basic concept and the first results of the Relay-MT algorithm. The presented algorithm does not always find the optimal translation. In the current approach, only the most recently generated translation is considered in subsequent hypothesis search steps. This algorithm could be extended by searching among a larger set of top hypotheses or even among previously rejected ones. The paper does not discuss the theoretical or numerical advantages of the algorithm, as research on it is still in its early phase. We have also not compared the presented method to other SOTA solutions for ensembling differently tokenized LLMs mentioned in Section 2.

After the official WMT26 submission deadline, we identified integrating Quality Estimation (QE) scores into the hypothesis reranking stage as a promising extension. This modification was not part of the submitted system, and its systematic evaluation is left for future work.

Acknowledgments

This paper was prepared with the assistance of AI tools. The use of Generative AI was limited to cases permitted under the *ACL Policy on Publication Ethics*. AI was used solely for related-work research, translations, language assistance and proof-reading.

Ethical Considerations

We took special care to thoroughly review the existing literature to avoid plagiarism. If any concept or solution had been previously published and we failed to mention it, this was unintentional. We have taken care in our experiments to use only models and sources that are generally available and free for research purposes. If we have infringed any intellectual property rights in this paper, it was unintentional.

References

Shanbo Cheng, Yu Bao, Qian Cao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, and 8 others. 2025. [Seed-x: Building strong multilingual translation llm with 7b parameters](#). *Preprint*, arXiv:2507.13618.

Adam Dobrowolski, Paweł Przewłocki, Paweł Przybysz, Marcin Szymański, and Dawid Siwicki. 2025. [A*](#)

[decoding for machine translation in LLMs - SRPOL participation in WMT2025](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 666–670, Suzhou, China. Association for Computational Linguistics.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. [TranslateGemma technical report](#). *Preprint*, arXiv:2601.09012.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.

Marcin Junczys-Dowmunt. 2019. [Dual conditional cross-entropy filtering of noisy parallel corpora](#). *Preprint*, arXiv:1809.00197.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [Metricx-24: The google submission to the wmt 2024 metrics shared task](#). *Preprint*, arXiv:2410.03983.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Hao Peng, Ximing Lu, Dragomir Radev, Yejin Choi, and Noah A. Smith. 2022. [Twist decoding: Diverse generators guide each other](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4909–4923, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Beirard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025a. [Command-a-translate: Raising the bar of machine translation with difficulty filtering](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 789–799, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Drach, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thammie Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025b. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale. In *Proceedings of the Eleventh Conference on Machine Translation*. 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, and 3 others. 2024. [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*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.
- Shankar Kumar and William Byrne. 2004. [Minimum Bayes-risk decoding for statistical machine translation](#). In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 169–176.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. [Eurollm: Multilingual language models for europe](#). *Preprint*, arXiv:2409.16235.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa 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*, pages 841–848, Singapore. Association for Computational Linguistics.
- Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeira, Amin Farajian, and André F. T. Martins. 2025. [Tower+: Bridging generality and translation specialization in multilingual llms](#). *Preprint*, arXiv:2506.17080.
- 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, Online. Association for Computational Linguistics.
- Yuzhe Shang, Pengzhi Gao, Wei Liu, Jian Luan, and Jinsong Su. 2026. [Scaling model and data for multilingual machine translation with open large language models](#). *Preprint*, arXiv:2602.11961.
- Shaomu Tan and Christof Monz. 2025. [ReMedy: Learning machine translation evaluation from human preferences with reward modeling](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4370–4387, Suzhou, China. Association for Computational Linguistics.
- Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and Andre Martins. 2024. [xTower: A multilingual LLM for explaining and correcting translation errors](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA. Association for Computational Linguistics.
- Giorgos Vernikos and Andrei Popescu-Belis. 2024. [Don’t rank, combine! combining machine translation hypotheses using quality estimation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12087–12105, Bangkok, Thailand. Association for Computational Linguistics.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. [Diverse beam search: Decoding diverse solutions from neural sequence models](#). *Preprint*, arXiv:1610.02424.
- Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. [Marco large translation model at WMT2025: Transforming translation capability in LLMs via quality-aware training and decoding](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 587–593, Suzhou, China. Association for Computational Linguistics.
- Rachel Wicks, Kartik Ravisankar, Xincheng Yang, Philipp Koehn, and Matt Post. 2025. [Token-level ensembling of models with different vocabularies](#). *Preprint*, arXiv:2502.21265.
- Yangyifan Xu, Jinliang Lu, and Jiajun Zhang. 2024. [Bridging the gap between different vocabularies for LLM ensemble](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7140–7152, Mexico City, Mexico. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, and Guanwei Zhang. 2026. [Hy-mt2: A family of fast, efficient and powerful multilingual translation models in the wild](#). *Preprint*, arXiv:2605.22064.

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. 2025a. [Hy-mt1.5 technical report](#). *Preprint*, arXiv:2512.24092.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025b. [Hunyuan-mt technical report](#). *Preprint*, arXiv:2509.05209.

A Results of experiments for 13 target languages

Below tables present QE scores of the WMT26 blindset translations for 13 main directions by different LLMs. MBR rows show results of sampling of 20 hypotheses with temperature=0.7 reranked with COMET-MBR or chrF-MBR.

Last four rows in each table show:

- Results of the log-likelihood reranking described in this paper.
- Results after 10 iterations of Relay-MT algorithm.
- Final submission after error correction.
- Results of Relay-MT for an ensemble with highest QE values (not submitted to WMT26).

Columns present referenceless scores for the following models:

- COMETKiwi-XL (Rei et al., 2023)
- MetricX-24-XL (Juraska et al., 2024)
- xCOMET-XL (Guerreiro et al., 2024)
- ReMedy-2B (Tan and Monz, 2025)

Values of MetricX-24 are negated, other QEs are multiplied by 100. Best results are presented in bold.

Table 3: Czech/English → German

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	65.2	-2.35	84.1	63.0
HY-MT1.5-7B	66.1	-1.91	85.2	75.0
MiLMMT-46-12B	66.1	-2.13	85.3	68.7
Qwen3.5-27B	61.8	-2.28	76.6	60.3
Seed-X-PPO-7B	66.0	-2.15	84.8	61.9
Tower-Plus-9B	64.8	-2.29	84.2	65.6
command-a-translate	66.3	-2.08	85.3	68.0
translatagemma-27b	66.6	-1.84	84.9	73.9
HY-MT1.5-7B (COMET-MBR)	66.7	-1.84	85.7	79.7
HY-MT2-30B-A3B (COMET-MBR)	66.8	-1.84	86.1	78.6
translatagemma-27b (COMET-MBR)	66.9	-1.80	85.4	80.3
MultiLLM (chrF-MBR)	66.9	-1.83	86.2	81.2
MultiLLM (COMET-MBR)	67.4	-1.75	86.4	81.6
Reranking	66.8	-1.86	86.0	80.7
Relay-MT	66.9	-1.84	86.3	75.9
Our Submission	67.0	-1.79	86.4	78.2
Best-QE Ensemble	67.4	-1.72	86.4	83.7

Table 4: English → Czech

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	66.7	-4.66	78.4	45.7
HY-MT1.5-7B	69.4	-3.83	82.0	62.2
MiLMMT-46-12B	69.0	-4.13	81.1	53.1
Qwen3.5-27B	64.8	-4.62	73.4	48.8
Seed-X-PPO-7B	68.1	-4.41	80.1	44.6
Tower-Plus-9B	67.4	-4.48	79.1	49.3
command-a-translate	69.3	-3.99	81.8	54.3
translatagemma-27b	70.7	-3.53	82.7	65.3
HY-MT1.5-7B (COMET-MBR)	70.7	-3.63	81.5	72.2
HY-MT2-30B-A3B (COMET-MBR)	70.3	-3.67	83.3	64.9
translatagemma-27b (COMET-MBR)	71.3	-3.36	83.6	69.8
MultiLLM (chrF-MBR)	70.4	-3.67	83.2	65.4
MultiLLM (COMET-MBR)	70.9	-3.66	81.2	78.1
Reranking	70.1	-3.72	83.0	64.4
Relay-MT	70.5	-3.63	83.5	63.9
Our Submission	70.6	-3.54	83.8	66.2
Best-QE Ensemble	71.4	-3.32	83.7	69.4

Table 5: English → Estonian

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	67.4	-5.76	80.2	-12.2
HY-MT1.5-7B	69.0	-4.99	83.6	0.6
MiLMMT-46-12B	62.7	-6.46	78.3	6.5
Qwen3.5-27B	60.0	-6.76	60.5	-2.9
Seed-X-PPO-7B	55.9	-8.75	65.3	-17.0
Tower-Plus-9B	37.0	-13.90	46.8	-58.5
command-a-translate	62.2	-7.09	74.7	-5.8
translatagemma-27b	70.2	-4.74	82.3	3.5
HY-MT1.5-7B (COMET-MBR)	70.0	-4.67	85.4	6.5
HY-MT2-30B-A3B (COMET-MBR)	70.9	-4.78	85.4	6.6
translatagemma-27b (COMET-MBR)	71.4	-4.47	84.4	12.0
MultiLLM (chrF-MBR)	69.9	-4.82	85.2	5.5
MultiLLM (COMET-MBR)	71.5	-4.36	86.9	10.3
Reranking	70.0	-4.86	85.0	4.5
Relay-MT	70.2	-4.74	85.4	2.1
Our Submission	70.4	-4.59	85.7	3.9
Best-QE Ensemble	71.4	-4.42	85.7	10.7

Table 6: English → Indonesian

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	48.0	-8.71	47.5	7.2
HY-MT1.5-7B	66.9	-3.17	79.0	68.3
MiLMMT-46-12B	65.2	-4.30	72.7	57.4
Qwen3.5-27B	65.7	-4.21	72.2	57.3
Seed-X-PPO-7B	64.0	-4.88	68.9	49.0
Tower-Plus-9B	56.1	-6.61	58.3	28.8
command-a-translate	65.4	-4.37	71.9	53.9
translatagemma-27b	67.6	-3.40	76.1	61.5
HY-MT1.5-7B (COMET-MBR)	67.2	-3.07	80.3	77.3
HY-MT2-30B-A3B (COMET-MBR)	68.0	-3.03	81.1	78.3
translatagemma-27b (COMET-MBR)	67.8	-3.27	77.1	74.5
MultiLLM (chrF-MBR)	67.5	-3.20	79.1	81.6
MultiLLM (COMET-MBR)	67.9	-2.87	81.3	79.3
Reranking	67.8	-3.08	80.9	67.4
Relay-MT	67.7	-3.07	80.8	67.7
Our Submission	67.9	-3.03	81.2	81.0
Best-QE Ensemble	67.9	-2.74	81.5	84.4

Table 7: English → Icelandic

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	28.5	-17.13	38.3	-82.6
HY-MT1.5-7B	68.3	-5.45	78.4	-34.7
MiLMMT-46-12B	58.0	-9.09	65.4	-25.9
Qwen3.5-27B	64.6	-7.07	72.3	-42.0
Seed-X-PPO-7B	59.1	-6.24	75.2	14.0
Tower-Plus-9B	69.4	-5.12	78.4	-48.2
command-a-translate	61.1	-8.48	67.3	-36.9
translategemma-27b	68.2	-5.21	76.9	-36.9
HY-MT1.5-7B (COMET-MBR)	69.7	-5.00	80.5	-26.8
HY-MT2-30B-A3B (COMET-MBR)	70.4	-4.81	81.5	-24.0
translategemma-27b (COMET-MBR)	70.0	-4.66	80.1	-22.4
MultiLLM (chrF-MBR)	70.0	-5.12	80.3	-26.1
MultiLLM (COMET-MBR)	71.2	-4.40	82.6	-19.9
Reranking	70.1	-4.99	80.3	-26.9
Relay-MT	70.2	-4.85	81.0	-26.7
Our Submission	70.1	-4.65	81.0	-27.8
Best-QE Ensemble	70.5	-4.30	81.8	-20.5

Table 10: English → Russian

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	67.0	-3.47	79.3	64.7
HY-MT1.5-7B	69.9	-2.14	84.1	81.9
MiLMMT-46-12B	69.3	-2.72	82.4	73.1
Qwen3.5-27B	66.8	-2.75	77.3	74.4
Seed-X-PPO-7B	69.3	-2.86	81.4	58.5
Tower-Plus-9B	68.3	-2.95	80.8	70.1
command-a-translate	69.2	-2.90	81.9	68.9
translategemma-27b	70.9	-2.11	83.4	83.4
HY-MT1.5-7B (COMET-MBR)	70.5	-2.00	84.9	84.1
HY-MT2-30B-A3B (COMET-MBR)	71.0	-2.00	85.1	84.0
translategemma-27b (COMET-MBR)	71.4	-2.01	84.1	86.0
MultiLLM (chrF-MBR)	71.0	-2.09	84.8	84.7
MultiLLM (COMET-MBR)	73.5	-1.54	84.9	97.1
Reranking	70.7	-2.06	84.9	84.6
Relay-MT	71.0	-2.02	85.3	84.1
Our Submission	71.2	-1.96	85.4	85.6
Best-QE Ensemble	71.4	-1.82	85.7	86.8

Table 8: Chinese Simplified/English → Japanese

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	61.0	-6.01	55.9	70.5
HY-MT1.5-7B	63.6	-4.96	63.1	81.6
MiLMMT-46-12B	62.7	-5.37	59.6	77.8
Qwen3.5-27B	63.8	-5.21	59.3	80.2
Seed-X-PPO-7B	61.5	-5.67	56.0	68.7
Tower-Plus-9B	62.4	-5.52	58.9	75.8
command-a-translate	62.8	-5.49	57.4	72.1
translategemma-27b	63.8	-4.80	60.6	81.1
HY-MT1.5-7B (COMET-MBR)	64.0	-4.82	64.7	89.7
HY-MT2-30B-A3B (COMET-MBR)	64.3	-4.78	64.7	88.8
translategemma-27b (COMET-MBR)	64.2	-4.64	61.9	90.6
MultiLLM (chrF-MBR)	64.0	-4.88	64.0	92.1
MultiLLM (COMET-MBR)	64.6	-4.65	65.3	90.9
Reranking	63.9	-4.91	63.9	81.1
Relay-MT	64.4	-4.83	64.6	81.7
Our Submission	64.4	-4.78	65.0	92.3
Best-QE Ensemble	64.3	-4.58	64.5	95.3

Table 11: English → Thai

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	26.9	-13.09	32.1	-46.0
HY-MT1.5-7B	65.0	-3.92	73.7	70.5
MiLMMT-46-12B	64.2	-4.47	69.1	62.4
Qwen3.5-27B	64.8	-4.41	68.5	68.4
Seed-X-PPO-7B	62.9	-5.05	62.7	46.7
Tower-Plus-9B	48.9	-8.11	43.5	22.2
command-a-translate	59.1	-5.95	55.7	47.5
translategemma-27b	65.6	-4.07	71.0	65.8
HY-MT1.5-7B (COMET-MBR)	65.4	-3.76	75.3	81.6
HY-MT2-30B-A3B (COMET-MBR)	65.8	-3.64	75.7	77.1
translategemma-27b (COMET-MBR)	65.8	-3.94	72.3	78.3
MultiLLM (chrF-MBR)	65.6	-3.85	74.0	83.3
MultiLLM (COMET-MBR)	65.7	-3.61	76.4	81.3
Reranking	65.5	-3.73	75.4	84.8
Relay-MT	65.7	-3.78	76.1	85.9
Our Submission	65.7	-3.72	76.2	81.9
Best-QE Ensemble	66.0	-3.43	76.9	86.5

Table 9: English → Korean

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	68.1	-3.95	77.1	36.7
HY-MT1.5-7B	71.2	-3.29	83.0	50.7
MiLMMT-46-12B	70.1	-3.49	80.4	44.2
Qwen3.5-27B	70.1	-3.66	79.4	41.8
Seed-X-PPO-7B	68.7	-3.93	76.8	35.4
Tower-Plus-9B	69.7	-3.71	79.2	42.4
command-a-translate	70.6	-3.60	80.6	44.2
translategemma-27b	70.4	-3.34	81.2	46.8
HY-MT1.5-7B (COMET-MBR)	71.5	-3.22	83.7	55.5
HY-MT2-30B-A3B (COMET-MBR)	71.4	-3.23	83.4	52.4
translategemma-27b (COMET-MBR)	70.7	-3.26	82.2	53.2
MultiLLM (chrF-MBR)	71.1	-3.26	83.1	54.8
MultiLLM (COMET-MBR)	71.9	-3.13	84.3	56.9
Reranking	71.4	-3.22	83.5	55.0
Relay-MT	71.4	-3.24	83.6	56.0
Our Submission	71.6	-3.20	83.7	56.3
Best-QE Ensemble	71.9	-3.02	84.2	57.1

Table 12: Czech/English → Ukrainian

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	59.9	-5.90	65.4	51.9
HY-MT1.5-7B	63.0	-4.59	72.2	66.5
MiLMMT-46-12B	60.5	-4.86	69.8	58.0
Qwen3.5-27B	62.6	-5.11	69.2	60.9
Seed-X-PPO-7B	62.3	-5.14	69.7	50.6
Tower-Plus-9B	61.7	-5.44	67.9	55.4
command-a-translate	62.5	-4.96	70.5	56.4
translategemma-27b	64.8	-4.25	71.6	65.0
HY-MT1.5-7B (COMET-MBR)	63.7	-4.37	73.6	71.0
HY-MT2-30B-A3B (COMET-MBR)	65.1	-4.22	74.6	72.7
translategemma-27b (COMET-MBR)	65.4	-4.09	72.8	73.1
MultiLLM (chrF-MBR)	64.6	-4.36	73.7	77.7
MultiLLM (COMET-MBR)	65.3	-4.04	75.4	75.4
Reranking	64.4	-4.34	73.6	65.2
Relay-MT	65.0	-4.28	74.5	77.3
Our Submission	65.0	-4.19	74.7	76.8
Best-QE Ensemble	65.2	-4.05	74.9	79.8

Table 13: Czech/English → Vietnamese

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	28.8	-12.32	30.5	-36.7
HY-MT1.5-7B	54.1	-4.76	49.6	18.1
MiLMMT-46-12B	51.0	-5.64	47.4	3.7
Qwen3.5-27B	52.1	-5.34	47.3	16.5
Seed-X-PPO-7B	50.9	-5.72	46.3	3.9
Tower-Plus-9B	38.8	-9.43	38.1	-10.1
command-a-translate	52.0	-5.40	48.6	6.7
translategemma-27b	53.2	-4.94	48.4	9.4
HY-MT1.5-7B (COMET-MBR)	54.8	-4.61	50.4	15.7
HY-MT2-30B-A3B (COMET-MBR)	54.7	-4.51	50.9	15.4
translategemma-27b (COMET-MBR)	53.9	-4.75	49.6	14.0
MultiLLM (chrF-MBR)	54.0	-4.73	50.5	20.5
MultiLLM (COMET-MBR)	55.1	-4.41	51.0	15.5
Reranking	54.3	-4.60	51.0	21.2
Relay-MT	54.9	-4.52	51.8	22.2
Our Submission	54.9	-4.32	52.0	19.9
Best-QE Ensemble	55.1	-4.19	52.3	21.4

Table 14: English → Chinese, Simplified

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	66.0	-3.29	68.7	67.2
HY-MT1.5-7B	70.5	-2.28	77.4	91.3
MiLMMT-46-12B	69.2	-2.73	73.1	81.0
Qwen3.5-27B	69.8	-2.62	74.1	86.1
Seed-X-PPO-7B	69.7	-2.48	73.9	83.5
Tower-Plus-9B	67.5	-2.97	71.1	75.5
command-a-translate	68.9	-2.88	72.2	75.6
translategemma-27b	71.0	-2.31	75.0	85.9
HY-MT1.5-7B (COMET-MBR)	70.9	-2.22	78.2	95.8
HY-MT2-30B-A3B (COMET-MBR)	71.1	-2.17	78.3	93.5
translategemma-27b (COMET-MBR)	71.3	-2.23	76.0	93.8
MultiLLM (chrF-MBR)	70.9	-2.27	77.5	95.9
MultiLLM (COMET-MBR)	71.6	-2.09	79.0	96.4
Reranking	71.0	-2.25	78.0	96.0
Relay-MT	71.2	-2.20	78.1	89.9
Our Submission	71.2	-2.16	78.2	96.4
Best-QE Ensemble	71.5	-2.07	78.6	97.4

Table 15: English → Chinese, Traditional Taiwan

Model	COMETKiwi	MetricX	xCOMET	ReMedy
EuroLLM-22B	60.5	-4.60	55.2	68.5
HY-MT1.5-7B	64.7	-3.41	65.3	85.9
MiLMMT-46-12B	63.9	-3.76	61.0	81.9
Qwen3.5-27B	65.3	-3.62	62.7	84.2
Seed-X-PPO-7B	63.1	-3.77	59.4	69.7
Tower-Plus-9B	62.5	-4.13	56.8	73.2
command-a-translate	63.9	-4.05	60.3	72.6
translategemma-27b	67.8	-3.31	62.3	80.3
HY-MT1.5-7B (COMET-MBR)	64.8	-3.29	66.3	94.9
HY-MT2-30B-A3B (COMET-MBR)	70.0	-2.84	67.9	84.2
translategemma-27b (COMET-MBR)	68.4	-3.15	64.4	94.9
MultiLLM (chrF-MBR)	65.9	-3.30	66.8	100.7
MultiLLM (COMET-MBR)	70.6	-3.22	65.0	106.7
Reranking	65.7	-3.24	67.7	99.8
Relay-MT	65.8	-3.23	68.4	101.3
Our Submission	65.8	-3.17	68.3	101.6
Best-QE Ensemble	66.4	-3.02	68.8	101.8

Table 16: List of translation models used in experiments

Model name	Hugging Face URL
Marco-MT-Algharb	https://huggingface.co/ATH-MaaS/Marco-MT-Algharb
EuroLLM-22B-Instruct	https://huggingface.co/utter-project/EuroLLM-22B-Instruct-2512
command-a-translate	https://huggingface.co/CohereLabs/command-a-translate-08-2025
Hunyuan-MT-7B	https://huggingface.co/tencent/Hunyuan-MT-7B
HY-MT1.5-7B	https://huggingface.co/tencent/HY-MT1.5-7B
Hy-MT2-30B-A3B	https://huggingface.co/tencent/Hy-MT2-30B-A3B
Hy-MT2-7B	https://huggingface.co/tencent/Hy-MT2-7B
MiLMMT-46-12B	https://huggingface.co/xiaomi-research/MiLMMT-46-12B-v0.1
Qwen3.5-27B	https://huggingface.co/Qwen/Qwen3.5-27B
Qwen3.5-122B-A10B	https://huggingface.co/Qwen/Qwen3.5-122B-A10B
Qwen3.6-27B	https://huggingface.co/Qwen/Qwen3.6-27B
Seed-X-PPO-7B	https://huggingface.co/ByteDance-Seed/Seed-X-PPO-7B
Tower-Plus-9B	https://huggingface.co/Unbabel/Tower-Plus-9B
translategemma-27b-it	https://huggingface.co/google/translategemma-27b-it