Laniqo at WMT25 General Translation Task: Self-Improved and Retrieval-Augmented Translation

Kamil Guttmann[1,2], **Zofia Rostek**[1], **Adrian Charkiewicz**[1,2], **Antoni Solarski**[1,†], **Mikołaj Pokrywka**[1,2], **Artur Nowakowski**[1,2]
[1] Laniqo, Poznań, Poland
[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
{name}.{surname}@laniqo.com

Abstract

This work describes Laniqo's submission to the constrained track of the WMT25 General MT Task. We participated in 11 translation directions. Our approach combines several techniques: fine-tuning the EuroLLM-9B-Instruct model using Contrastive Preference Optimization on a synthetic dataset, applying Retrieval-Augmented Translation with human-translated data, implementing Quality-Aware Decoding, and performing postprocessing of translations with a rule-based algorithm. We analyze the contribution of each method and report improvements at every stage of our pipeline.

1 Introduction

In this paper, we describe Laniqo's submission to the WMT 2025 General MT Task. We participated in the constrained track of the shared task, which limited our system to a maximum of 20 billion total parameters. This year's task focused on the translation of document-level data sampled from four domains: news, social, literary, and speech. Additionally, the provided testset contained metadata from different modalities, i.e. screenshots of posts from social media and audio recordings for the speech domain. Furthermore, each testset entry contained a domain-specific prompt for Large Language Models (LLMs).

We based our system on the `EuroLLM-9B-Instruct`[1] (Martins et al., 2025) model. We participated in all of the translation directions supported by the model, resulting in the following 11 distinct directions: Czech to German; Czech to Ukrainian; English to Chinese; English to Czech; English to Estonian; English to Italian; English to Japanese; English to Korean; English to Russian; English to Ukrainian; and Japanese to Chinese.

While the base model provides strong multilingual translation capabilities, to further improve translation quality across multiple language pairs, we developed a multi-stage translation pipeline consisting of the following methods:

1. **Contrastive Preference Optimization**

   We fine-tuned the model using Contrastive Preference Optimization (CPO) (Xu et al., 2024) on a synthetically created preference dataset covering eight language pairs. The model weights can be accessed via Hugging Face[2].

2. **Retrieval-Augmented Translation** To bring machine translation outputs closer to human-level quality, we incorporated Retrieval-Augmented Translation (RAT) into our pipeline. This component retrieves semantically similar segments from a pre-indexed vector database and dynamically integrates them as few-shot examples in the model prompt.

3. **Quality-Aware Decoding** First, we generated multiple candidates for each sentence, and then we applied a reranking process. We scored each candidate using a reference-free quality estimation metric, identifying translations that are likely to be of high quality, to reduce the number of candidates. This was followed by Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004) to select translations with the lowest expected quality loss across the sampled hypotheses.

4. **Postprocessing**

   The final translation is further refined through rule-based postprocessing, which includes restoration of URLs and emojis, preservation of original casing, and normalization of language-specific quotation marks.

[†] Work done while working at Laniqo

[1] https://huggingface.co/utter-project/EuroLLM-9B-Instruct

[2] https://huggingface.co/laniqo/WMT25-EuroLLM-9B-CPO

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 778–788
November 8–9, 2025.

2 Related Work

LLMs have become the dominant approach in the field, overtaking smaller Neural Machine Translation (NMT) models (Kocmi et al., 2024), particularly following the release of open-source, multilingual LLMs, such as EuroLLM and Tower+ (Rei et al., 2025). A fundamental advantage of LLMs is their ability to process instructions provided directly within the prompt.

Studies have shown that few-shot prompting outperforms zero-shot translation and that selecting examples with high lexical similarity, employing methods such as fuzzy matching, can further enhance translation quality (Moslem et al., 2023).

Quality-Aware Decoding (QAD) (Fernandes et al., 2022) is an established method for improving translation quality. It facilitates MBR decoding, which uses a translation quality metric as the scoring function to rerank a list of translation candidates. Subsequent research has consistently demonstrated the effectiveness of this method in improving translation outputs (Nowakowski et al., 2022; Rei et al., 2024).

MBR decoding is computationally expensive because it requires generating numerous translation candidates and making pairwise comparisons between them. The computational cost, apparent during inference, can be reduced through MBR self-improvement. (Guttmann et al., 2024; Finkelstein and Freitag, 2024), a technique that involves fine-tuning a model using outputs selected by MBR. The self-improvement process can be framed as a preference learning task. Methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) and CPO have been shown to be more effective than Supervised Fine-Tuning (SFT) for such tasks. Hence, using these preference optimization methods with MBR self-improvement has been shown to yield further enhancements in terms of translation quality (Yang et al., 2024).

NMT and LLMs often struggle with specific token types, such as numbers, URLs, or emojis (Wisniewski et al., 2025a). These models may incorrectly translate or even omit such tokens. While these errors are critical for the user, they are often not captured by neural metrics. Therefore, a simple post-processing step can become highly valuable. By employing a straightforward rule-based method as proposed in previous work (Nowakowski et al., 2022; Wu et al., 2024), these errors can be detected and corrected.

3 Approach

3.1 Data

We created a synthetic preference dataset, covering eight language pairs, namely English to Arabic, Korean, Japanese, Ukrainian, Czech, Chinese, Russian, and Estonian. We excluded Italian from the target languages due to its late inclusion in the human-evaluated languages of the WMT25 General MT shared task. To construct the dataset, we sampled 10,000 English document-level examples from the NewsPaLM corpus (Finkelstein et al., 2024). For each source example, we generated 64 translation candidates with `EuroLLM-9B-Instruct` using epsilon sampling with $\epsilon = 0.02$ and $T = 1$, following previous work (Freitag et al., 2023). Then we reranked the candidate list using MBR decoding, with `wmt22-comet-da`[3] (Rei et al., 2022) serving as the utility metric. From the reranked candidate list, we selected the 1st, 32nd, and 64th translations to form the chosen, medium, and rejected examples, respectively, following the BMW strategy (Yang et al., 2024).

3.2 CPO

We used the dataset described above to align the `EuroLLM-9B-Instruct` model-generated outputs more closely with neural MT quality metrics, which are known to correlate highly with human preferences. To achieve this, we applied CPO, implementing the fine-tuning in a parameter-efficient manner using Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023).

We trained the QLoRA on a single A100 GPU using the Unsloth[4] framework. The specific training hyperparameters are detailed in Table 1. Although training continued beyond step 2,000, evaluation on the WMT24++ (Deutsch et al., 2025) testset indicated that the checkpoint corresponding to approximately 1.32 epochs over the entire dataset achieved the highest score under the COMET metric.

In preliminary evaluations, we observed that the English to Arabic translation direction yielded notably low evaluation scores. This issue may be the result of EuroLLM's support for Modern Standard Arabic, and not the Egyptian dialect that is evaluated during WMT25. Consequently, this pair was

[3] https://huggingface.co/Unbabel/wmt22-comet-da

[4] https://unsloth.ai/

Parameter Category	Value/Description
QLoRA Configuration	
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.0
CPO Objective Configuration	
Loss Type	Sigmoid
Beta (β)	0.7
Label Smoothing	0.15
CPO Alpha (α)	1.0
General Training Configuration	
Per-Device Batch Size	4
Gradient Accumulation Steps	12
Effective Global Batch Size	48
Learning Rate	5.0×10^{-7}
LR Scheduler Type	Cosine
Warm-up Steps	100

Table 1: CPO and QLoRA Configuration Parameters

excluded from subsequent experiments.

3.3 Retrieval-Augmented Translation

To enhance the translation quality and adaptability of our machine translation system, we implemented a dynamic few-shot example selection mechanism. The objective of this approach is to provide semantically relevant human-translated examples within the translation prompt in order to guide the model towards more accurate and fluent translations across diverse input styles and domains. This is obtained by applying a Retrieval-Augmented Generation pipeline for the translation task (Retrieval-Augmented Translation).

For each given input segment, we retrieved a set of few-shot examples from a vector database constructed by indexing high-quality, human-translated data from previous WMT testsets[5], WMT24++, FLORES-200 (NLLB Team et al., 2022), and NTREX-128 (Federmann et al., 2022), covering all language pairs supported by our system. We selected these datasets specifically for their established reputation within the machine translation community as sources of high-quality, human-translated examples, covering multiple domains.

We used the Qdrant[6] vector database to efficiently store and retrieve similar examples. We calculated embeddings for all segments in this database using the `e5-multilingual-base`[7] (Wang et al., 2024) model, which was selected for its strong performance in multilingual semantic similarity tasks. To identify the most semantically similar examples, we used cosine similarity to compare the embedding of the currently translated source segment against the database entries. The top three closest entries are then retrieved, along with their translations, and used as few-shot examples.

The experimental setting described above was determined by preliminary experiments conducted on the WMT24++ testset. To achieve unbiased results, we excluded the testset from the vector database. These ablation studies evaluated three primary factors: the choice of embedding model, the number of few-shot examples, and the examples' semantic similarity to the source sentence. We compared two models for generating embeddings: `multilingual-e5-base` and `EuroLLM-9B-Instruct`, and tested performance when providing top-k examples for $k \in 1, 3, 5$, optionally using a similarity score threshold of 0.8. The detailed results of these experiments are presented in Table 2.

3.4 Quality-Aware Decoding

For the final translation step, we build upon the QLoRA and RAT pipeline, and integrate them with QAD (Fernandes et al., 2022), which we achieve through QE reranking and MBR decoding, to further enhance the translation quality. Due to the model's limited context window and the inclusion of few-shot examples in the prompt, we split the WMT25 testset using newline characters rather than paragraphs, as the latter often produced input segments that exceeded the model's context window limit. For each source segment, 128 translation candidates are generated through epsilon sampling with identical parameters to those used during the creation of the preference dataset. The candidate pool is then pruned to eight candidates per source segment through a QE reranking process, utilizing `wmt23-cometkiwi-da-xl`[8] (Rei et al., 2023) as the underlying scoring function. Finally, MBR decoding, with `xCOMET-XL`[9] (Guerreiro et al., 2024) as

[5] https://data.statmt.org/wmt24/general-mt/wmt24_GeneralMT-devsets.zip

[6] https://qdrant.tech/

[7] https://huggingface.co/intfloat/multilingual-e5-base

[8] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

[9] https://huggingface.co/Unbabel/XCOMET-XL

Embeddings	Top-k	Threshold	xCOMET ↑	xCOMET-QE ↑	BLEU ↑
–	–	–	0.7909	0.7846	26.58
multilingual-e5-base	1	–	0.7935	0.7857	26.76
	3	–	**0.7955**	**0.7875**	26.81
	3	0.8	0.7949	0.7872	26.60
	5	0.8	0.7954	**0.7875**	26.74
EuroLLM	1	–	0.7931	0.7852	**26.82**
	3	–	0.7946	0.7868	26.70
	3	0.8	0.7937	0.7860	26.70
	5	0.8	0.7942	0.7861	26.77

Table 2: Performance comparison of different Retrieval-Augmented Translation approaches on the WMT24++ testset. The reported scores are macro-averages calculated across all language pairs that we participated in, excluding the English to Italian translation direction.

the utility function, is applied to select the final translation.

3.5 Postprocessing

We applied a series of post-processing steps to our system's translations to further refine their quality and ensure adherence to language-specific requirements:

- **Casing Restoration:** To maintain typographical consistency, we applied the corresponding casing to the target translation if the source segment was entirely in uppercase, lowercase, or titlecase.
- **Quotation Mark Normalization:** We replaced generic double quotation marks (") in the target outputs with their correct language-specific forms to align with punctuation standards. For example, we converted them to forms such as Chinese (“ ”), Czech („ “), Estonian („ “), Italian (« »), Japanese (「 」), Korean („ “), Russian (« »), and Ukrainian (« »).
- **URL Restoration:** To preserve correct external links, we replaced any URL identified in the target translation that differed from its source counterpart with the exact URL from the source, thereby preventing any discrepancies between the source sentence and the translation.
- **Emoji Restoration:** To ensure accurate emoji representation, we corrected discrepancies between source and target emojis. If a single emoji appeared in both the source and target but differed, the target's emoji was replaced with the source's. Furthermore, any sequences of emojis located at the beginning or end of the source segment were compared with those in the target, and discrepancies led to the replacement of the target's boundary emojis with the source's.

3.6 Discarded Experimental Approaches

We also conducted several additional experiments that were excluded from our final submission due to inconsistent or negative results.

These included applying Named Entity Recognition (NER) to improve handling and transferring named entities during translation. We also investigated grammar correction for texts from the speech domain, motivated by the assumption that such texts might contain specific grammatical errors introduced during Automatic Speech Recognition. Furthermore, we tested the use of domain-specific prompts to better adapt the system to particular content areas. All of the above experiments resulted in negative outcomes according to automatic evaluation metrics, and therefore were not pursued further.

Detailed descriptions of these experiments are provided in Appendix A (Named-Entity Recognition), Appendix B (Grammar Correction), and Appendix C (Domain-Specific Prompt).

4 Results

We evaluated our system with the `xCOMET-XL`, `ReMedy-9B-24`[10] (Tan and Monz, 2025), and `MetricX-24-Hybrid-XL`[11] (Juraska et al., 2024) automatic translation evaluation metrics. Due to

[10] https://huggingface.co/ShaomuTan/ReMedy-9B-24

[11] https://huggingface.co/google/metricx-24-hybrid-xl-v2p6

System	xCOMET-QE ↑	ReMedy-QE ↑	MetricX-QE ↓
WMT25 testset prompt	0.7345	0.6298	10.5838
Baseline	0.7391	0.6322 *	10.4312
+CPO	0.7537 *	0.6405 *	9.5361 *
+RAT	0.7414	0.6334	10.1587
+CPO +RAT	0.7560	0.6411	9.5319
+CPO +RAT +QAD	**0.8343** *	**0.6435** *	9.2849
+CPO +RAT +QAD +postprocessing	0.8339	0.6431	**9.2810**

Table 3: Macro average system quality. Results of xCOMET-QE, ReMedy-QE and MetricX-QE automatic evaluation metrics on the concatenated WMT25 testset for Czech → German, Ukrainian; English → Czech, Estonian, Italian, Japanese, Korean, Russian, Ukrainian, Chinese; Japanese → Chinese (general collection only). Results marked with an asterisk (*) are statistically significant compared to the previous pipeline step results; the baseline was compared with the WMT25 testset prompt solution.

```
<|im_start|>system
You are a professional {src_lang} to {tgt_lang} translator.
Your goal is to accurately convey the meaning and nuances of the
original {src_lang} text while adhering to {tgt_lang} grammar,
vocabulary, and cultural sensitivities.
<|im_end|>
<|im_start|>user
Translate the following {src_lang} source text to {tgt_lang}:
{src_lang}: {source}
{tgt_lang}: <|im_end|>
<|im_start|>assistant
```

Listing 1: Baseline translation prompt.

the lack of access to reference translations, the scores were calculated in quality estimation (QE) mode, based on source texts and hypotheses only.

The results presented in Table 3 show the improvements achieved after each translation pipeline step. Initially, we employed greedy decoding and the prompts provided in the WMT25 testset to compare the results with our baseline translation prompt presented in Listing 1. Based on these results, we decided to use our prompt in further experiments.

Although the use of RAT alone does not visibly enhance quality, its combination with CPO results in substantially greater gains. Overall, these findings suggest that fine-tuning through CPO effectively enhanced translation quality, aligning the model's outputs more closely with human preferences as indicated by quality estimation metrics.

QAD yields the most significant improvements across the entire processing pipeline. We specifically noted improvements in the `xCOMET-XL` scores. However, it is important to consider that, due to MINT (Pombal et al., 2025), `xCOMET-XL` as an interfering metric may be biased and can't be used to evaluate the model fairly. For this reason, we also present results from other neural metrics, providing a more comprehensive assessment of translation quality.

Additionally, rule-based postprocessing helps to avoid translation errors, even though these improvements are not reflected in the evaluation metrics due to their limitations.

Moreover, we performed statistical tests using the Paired Bootstrap Resampling method (Koehn, 2004). We sampled $s = 1000$ times with $n = 0.4 * testset\_length$ segments and p-value $p = 0.05$. We compared the results of each pipeline step with the previous one, and the baseline to the WMT25 testset prompt solution. The results show that the baseline increase in the `ReMedy-QE` score is statistically significant compared to the results of the WMT25 testset prompts. Furthermore, CPO is one of the most meaningful steps in the entire pipeline, showing a significant difference compared to the baseline according to all the considered metrics. While adding the RAT step improves the results slightly, using the QAD method is the second most

Source language	Target language	xCOMET-QE ↑	ReMedy-QE ↑	MetricX-QE ↓
Czech	German	0.9359	0.6400	5.7178
	Ukrainian	0.9239	0.6398	8.0289
English	Czech	0.9016	0.6529	10.3234
	Estonian	0.8278	0.6470	11.7945
	Italian	0.8356	0.6528	9.5376
	Japanese	0.7831	0.6447	10.0123
	Korean	0.7884	0.6479	9.6596
	Russian	0.8426	0.6321	10.2086
	Ukrainian	0.8198	0.6413	10.5458
	Chinese	0.7410	0.6481	8.7743
Japanese	Chinese	0.7733	0.6272	7.4879
Macro average		0.8339	0.6431	9.2810

Table 4: System quality per language pair. Results of xCOMET-QE, ReMedy-QE and MetricX-QE automatic evaluation metrics on the concatenated WMT25 testset (general collection only).

important step, which significantly improves the results according to the `xCOMET-QE` and `ReMedy-QE` metrics. The differences in quality after the postprocessing step, including the decrease in two metrics, are not statistically significant.

Table 4 presents the results obtained for each language pair separately and the macro average score for the final translations.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects.

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.

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.

Mara Finkelstein and Markus Freitag. 2024. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. In *The Twelfth International Conference on Learning Representations*.

Mara Finkelstein, David Vilar, and Markus Freitag. 2024. Introducing the NewsPaLM MBR and QE dataset: LLM-generated high-quality parallel data outperforms traditional web-crawled data. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1355–1372, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. 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.

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.

Kamil Guttmann, Mikołaj Pokrywka, Adrian Charkiewicz, and Artur Nowakowski. 2024. Chasing COMET: Leveraging minimum Bayes risk decoding for self-improving machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 80–99, Sheffield, UK. European Association for Machine Translation (EAMT).

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. 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, Steinthór Steingrímsson, and Vilém Zouhar. 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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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 North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025. Adding chocolate to mint: Mitigating metric interference in machine translation.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, 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, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Shaomu Tan and Christof Monz. 2025. Remedy: Learning machine translation evaluation from human preferences with reward modeling.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer,

Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Dawid Wisniewski, Mikolaj Pokrywka, and Zofia Rostek. 2025a. Do not change me: On transferring entities without modification in neural machine translation – a multilingual perspective.

Dawid Wisniewski, Antoni Solarski, and Artur Nowakowski. 2025b. Exploring the feasibility of multilingual grammatical error correction with a single llm up to 9b parameters: A comparative study of 17 models.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin Guo, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Ning Xie, and Hao Yang. 2024. Choose the final translation from NMT and LLM hypotheses using MBR decoding: HW-TSC's submission to the WMT24 general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 155–164, Miami, Florida, USA. Association for Computational Linguistics.

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

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

A Named-Entity Recognition

We explored the integration of Named-Entity Recognition (NER) into the translation pipeline, applying it in two distinct ways: (1) as injection into the prompt and (2) as glossary constraints.

Prompt Augmentation with NER. For each sentence, named entities were extracted using a multilingual NER model – `gliner_large-v2.5`[12]. These entities were then directly added to the prompt. The goal was to guide the model to pay closer attention to those terms during translation. For example, the prompt was extended to include an additional instruction such as: `The following named entities appear in the source text and should be preserved or accurately translated: {entity_list}`. Ideally, this mechanism could have encouraged accurate adaptation of names, locations, organizations, and other entities, but this approach yielded only negative results, in comparison to the baseline (Table 5).

NER as Terminology Constraints. As for the second approach, we treated named entities as terminology constraints. After extraction, the entities were translated individually to create source-target term pairs, including additional information from the NER model. These pairs were then injected into the prompt in a structured format for translating full sentences. The system translated each sentence independently using the input text along with the dictionary of domain-specific terminology pairs. Prior to translation, source-language named entities were replaced with their corresponding target-language equivalents. This was combined with explicit prompts designed to guide the model in retaining or correctly adapting the inserted terms. This resembled translation with terminology constraints but was adapted for automatically detected entities. This approach also failed to yield performance gains (Table 6), leading us to abandon the use of NER in this form.

[12] https://huggingface.co/gliner-community/gliner_large-v2.5

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	-0.0026	-0.01	0.50
	Ukrainian	-0.0080	-0.45	-0.73
English	Czech	-0.0031	-0.05	-0.06
	Estonian	-0.0044	-0.13	-0.21
	Japanese	-0.0039	-2.52	-0.44
	Korean	-0.0036	0.15	-0.32
	Russian	-0.0037	-0.44	-0.34
	Ukrainian	-0.0054	-0.34	-0.50
	Chinese	-0.0033	-2.23	-0.50
Japanese	Chinese	-0.0014	-0.33	0.20
Macro average		-0.0039	-0.64	-0.24

Table 5: The difference in automatic metrics between the NER-enhanced system and the baseline calculated on the WMT24++ testset.

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	-0.0099	-0.88	-1.17
	Ukrainian	-0.0063	-1.80	-1.51
English	Czech	-0.0130	-1.01	-0.84
	Estonian	-0.0238	-3.57	-3.45
	Japanese	-0.0074	-11.42	-2.13
	Korean	-0.0206	-3.77	-2.84
	Russian	-0.0231	-1.66	-1.87
	Ukrainian	-0.0172	-1.92	-2.18
	Chinese	-0.0180	-8.81	-2.88
Japanese	Chinese	-0.0017	-4.52	-1.10
Macro average		-0.0141	-3.94	-2.00

Table 6: The difference in automatic metrics between the system using NERs as terminology and the baseline calculated on the WMT24++ testset.

B Grammar Correction

Motivated by the hypothesis that speech domain texts may contain errors, we applied grammatical error correction to improve their quality prior to translation. Two approaches were attempted: (1) the utilization of the `Gemma` (Team et al., 2024) model for the purpose of grammatical correction prior to translation, and (2) the incorporation of additional information to the prompt employed for the correction of the text before translation.

Utilization of the Gemma model. The first approach involved using the `Gemma` model in the first step to perform grammatical correction, and then in the second step, using these corrected texts for standard translation with the `EuroLLM` model. We tested two model versions: `gemma-2-9b-it`[13] and `gemma-3-4b-it`[14], and several prompts to achieve the best possible results.

The best translations were obtained using a prompt described by Wisniewski et al. (2025b) and `gemma-3-4b-it` model, although this still resulted in a decrease in quality compared to the baseline approach, as shown in Table 7.

Grammar correction combined with translation. The second approach involved applying the same instructions used for the `Gemma` model directly to the translation prompt. The term `translator` was changed to `translator with correction capabilities`, and the following instruction was added: `Edit the following source text for spelling and grammar errors, make minimal changes, and use only the corrected text for translation. If the source text is already correct, translate it without any previous changes.`

The results of this experiment are presented in

[13] https://huggingface.co/google/gemma-2-9b-it
[14] https://huggingface.co/google/gemma-3-4b-it

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	-0.0078	-1.46	-1.15
	Ukrainian	-0.0100	-0.79	-0.94
English	Czech	-0.0013	-1.98	-1.03
	Estonian	-0.0013	-1.96	-0.96
	Japanese	0.0034	0.17	0.11
	Korean	0.0013	-0.54	-0.91
	Russian	0.0004	-1.20	-0.81
	Ukrainian	0.0078	-0.50	0.18
	Chinese	0.0016	-0.26	-0.30
Japanese	Chinese	0.0018	-0.94	-0.81
Macro average		-0.0004	-0.95	-0.66

Table 7: The difference in translation quality between the solution with grammar correction and the baseline solution on the WMT24++ testset for the speech domain data. A noticeable decline in translation quality is observed.

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	0.0048	0.54	0.68
	Ukrainian	-0.0016	-0.42	0.06
English	Czech	-0.0014	-0.31	-0.20
	Estonian	0.0019	-0.05	-0.11
	Japanese	0.0023	-0.06	0.06
	Korean	-0.0004	0.08	-0.29
	Russian	0.0034	0.10	-0.10
	Ukrainian	-0.0008	0.93	0.55
	Chinese	-0.0009	-0.26	-0.16
Japanese	Chinese	0.0019	-0.34	-0.32
Macro average		0.0009	0.02	0.02

Table 8: The difference in translation quality between the solution with extended translation prompt and the baseline solution on the WMT24++ testset for the speech domain data.

Table 8. Although the average outcomes slightly improved translation quality, these differences were not significant and varied between language pairs. Ultimately, this method was not employed in the final solution.

C Domain-Specific Prompt

Another approach involved using the domain information available in the dataset. We tested adding it to the translation prompt: `You are a professional {src_lang} to {tgt_lang} translator, specialized in the {domain} domain [...] Make sure to use vocabulary and grammatical structures appropriate for the {domain} domain. [...] Translate the following {domain} domain, {src_lang} source text to {tgt_lang}: [...]`. The results of this approach are presented in Table 9. This approach was not used in the final solution because it did not improve translation quality.

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	-0.0050	-0.31	-0.86
	Ukrainian	-0.0026	-0.39	-0.28
English	Czech	-0.0042	0.30	0.12
	Estonian	-0.0007	0.36	0.22
	Japanese	0.0003	1.57	0.51
	Korean	-0.0022	0.07	-0.20
	Russian	-0.0017	-1.00	-0.22
	Ukrainian	-0.0024	0.36	0.09
	Chinese	-0.0010	-0.01	0.03
Japanese	Chinese	-0.0031	-0.10	-0.16
Macro average		-0.0023	0.09	-0.07

Table 9: The difference in translation quality between the solution with a domain-specific prompt and the baseline solution on the WMT24++ testset.