Laniqo at WMT25 Terminology Translation Task: A Multi-Objective Reranking Strategy for Terminology-Aware Translation via Pareto-Optimal Decoding

Kamil Guttmann[1,2], **Adrian Charkiewicz**[1,2], **Zofia Rostek**[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 paper describes the Laniqo system submitted to the WMT25 Terminology Translation Task. Our approach uses a Large Language Model fine-tuned on parallel data augmented with source-side terminology constraints. To select the final translation from a set of generated candidates, we introduce Pareto-Optimal Decoding – a multi-objective reranking strategy. This method balances translation quality with term accuracy by leveraging several quality estimation metrics alongside Term Success Rate (TSR). Our system achieves TSR greater than 0.99 across all language pairs on the Shared Task testset, demonstrating the effectiveness of the proposed approach.

1 Introduction

The shared task consists of two tracks: sentence-level translation and document-level translation. Our submission focuses exclusively on the sentence-level track, where each sentence is translated independently using provided terminology constraints.

The shared task requires systems to be evaluated in three distinct modes:

- **No Terminology (`noterm`)**: The system is only provided with the input sentences.
- **Proper Terminology (`proper`)**: The system receives the input text along with a dictionary of domain-specific terminology pairs.
- **Random Terminology (`random`)**: The system is provided with the input text and a dictionary of randomly sampled terms from the source and target texts.

Our system builds upon the `EuroLLM-9B-Instruct` model[1] (Martins et al., 2025), which served as the baseline for our experiments in Terminology-Aware Machine Translation (MT). While this model already offers multilingual translation capabilities, it is not explicitly designed to handle translation with domain-specific terminology constraints. To further improve terminology control and translation quality, we explored several complementary strategies:

1. **Source-side terminology replacement**: Before translation, we replaced source-language terms 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.
2. **Fine-tuning on glossary-augmented data**: We fine-tuned the base model on parallel data augmented with terminology automatically aligned between source and target segments. The objective of this training was to expose the model to code-switched source sentences, enabling it to learn the mechanism for incorporating provided target-language terms during translation.
3. **Pareto-Optimal Decoding**: We propose a reranking strategy that integrates multiple reference-free Quality Estimation (QE) metrics along with terminology accuracy to select the most accurate and terminology-compliant translation candidate.

In addition, we investigated prompt engineering techniques, such as structured instructions and two-shot examples, aimed at improving system robustness.

The methods described above proved effective in both improving the correct use of terminology and maintaining overall translation quality, as confirmed by automatic metrics.

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

Proceedings of the Tenth Conference on Machine Translation, pages 1276–1283
November 8-9, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Handling specialized terminology in Neural Machine Translation (NMT) has received considerable research interest in recent years. The proposed methods can be broadly classified into three main categories:

- **Constrained decoding**: These methods modify the beam search algorithm by restricting the search space to ensure that the generated hypotheses include the specified terminology (Hokamp and Liu, 2017; Nowakowski and Jassem, 2021). Furthermore, the application of negative constraints has been studied for re-translating sentences where an initial translation failed to incorporate the required terms (Bogoychev and Chen, 2023).

- **Placeholding**: This approach involves replacing source terms with special placeholder tokens (Michon et al., 2020). The model is then trained to copy these placeholders into the hypothesis, enabling the target terms to be injected during post-processing. The main disadvantage of this method is that masking the source terms can result in a lack of context, which can lead to a degradation in fluency of the final translation.

- **Source Text Constraints**: This technique involves augmenting the training data with inline terminology constraints (Dinu et al., 2019; Bergmanis and Pinnis, 2021). The source text is augmented by adding the target term alongside its equivalent in the source language. These terms are typically annotated with source factors (Sennrich and Haddow, 2016) or pre-defined tags. This approach has proven particularly effective for morphologically rich languages when combined with Target Lemma Annotations (Bergmanis and Pinnis, 2021). The model generates the appropriate morphological form of the target term, ensuring that it adheres to the grammatical rules of the target language.

The emergence of multilingual Large Language Models (LLMs), such as EuroLLM (Martins et al., 2025) and Tower+ (Rei et al., 2025), represents a significant shift in MT towards the LLM era (Kocmi et al., 2024). The ability of LLMs to follow natural language instructions embedded within a prompt opens up new possibilities for adhering to terminology constraints.

Several possible strategies that leverage this capability have been explored. One approach involves using LLMs to generate term-rich synthetic data for fine-tuning traditional NMT models (Moslem et al., 2023b). Another line of work uses LLMs for automatic post-editing, prompting the model to inject missing terms into an existing translation (Bogoychev and Chen, 2023). A third, more direct method involves providing the LLM with the source text and a list of terminology constraints which must be included in the output (Moslem et al., 2023a).

Our approach builds on this direct prompting method by extending it through the integration of constraints in the source text.

While the primary objective of the Shared Task is to ensure the correct translation of specified terminology, maintaining the overall quality of the translation remains a critical factor. Well-established methods for improving the quality of machine translation, such as Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004) and QE reranking, have consistently demonstrated their effectiveness in various research studies in recent years (Nowakowski et al., 2022; Finkelstein and Freitag, 2024; Guttmann et al., 2024). However, these approaches are typically designed to optimize for a single metric.

This task requires the simultaneous optimization of two distinct aspects: term accuracy and general translation quality. To address this multi-objective problem, we propose a method that balances these potentially conflicting criteria.

3 Approach

3.1 Replace Method

Given a source segment and a glossary containing terminology pairs, we replaced each source term in the input text with its corresponding target-language term (see Table 1 for example). In contrast to previous works (Nieminen, 2023; Ri et al., 2021), where target terminology was appended to or replaced within the source text and subsequently enclosed by special tags, our approach, similar to the data augmentation method proposed by Song et al. (2019), directly replaces terms with their target-language equivalents without introducing any additional markup. Directly inserting target

```
<|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
Some words have been pre-translated. You may need to correct them
in the final translation for a better fit into the context.
{src_term1} -> {tgt_term1}
{src_term2} -> {tgt_term2}
{src_term3} -> {tgt_term3}

Translate the following {src_lang} source text to {tgt_lang}:
{src_lang}: {src_text_replaced}
{tgt_lang}: <|im_end|>
<|im_start|>assistant
```

Listing 1: Baseline terminology-aware translation prompt.

language terms into the source sentence results in code-switching, enabling the model to adapt the grammatical form of the target terms to fit them into the sentence structure during inference.

Source	"In the Switch **Data Provider** dialog:"
Terminology	{"data provider": "Daten-provider"}
Replaced	"In the Switch **Datenprovider** dialog:"

Table 1: Example of terminology handling method applied to the source sentence.

The replacement process involved lemmatizing both the source text and the glossary terms using the `simplemma` library (Barbaresi, 2021). Each source term was matched against the lemmatized input, and when a match was found, it was substituted with the corresponding target term.

3.2 Prompt

The base prompt was designed to ensure the integration of glossary terms. Since replacing source terms with their target-language equivalents can obscure grammatical cues needed for correct inflection, the prompt also includes a list of original source phrases alongside their corresponding target terms. This provides the model with additional context, helping it resolve potential ambiguities and adjust terminology to the surrounding syntax when necessary. The full prompt is shown in Listing 1.

During the experiments, we observed that a large proportion of the mistranslated terminology consisted of phrases that were identical or very similar in both the source and target languages - differing only in casing. When the model received a text in which a term had been replaced with an identical or nearly identical term in the target language, it did not recognize the change and consequently translated it into an equivalent term in the target language. We conducted additional experiments, making some adjustments to the prompt to draw the model's attention to this issue. As a result, we added the following instruction to the prompt: `(keep already translated {tgt_lang} words - {tgt_terms})`, for instance `Translate the following English source text to Spanish (keep already translated Spanish words - desea, Decida)`. This solution improved translation quality, leading to more frequent and correct usage of the specified terminology.

3.3 Few-Shot Prompting

We also conducted experiments on few-shot prompting. We found that providing two translation examples with terminology (in the appropriate source and target languages) improves translation quality, both in general translation metrics and in term accuracy, averaged across all three language pairs.

Parameter	Value/Description
LoRA Configuration	
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.15
General Training Configuration	
Per-Device Batch Size	4
Gradient Accumulation Steps	8
Learning Rate	1×10^{-4}
LR Scheduler Type	Inverse Square Root

Table 2: Fine-tuning hyperparameters

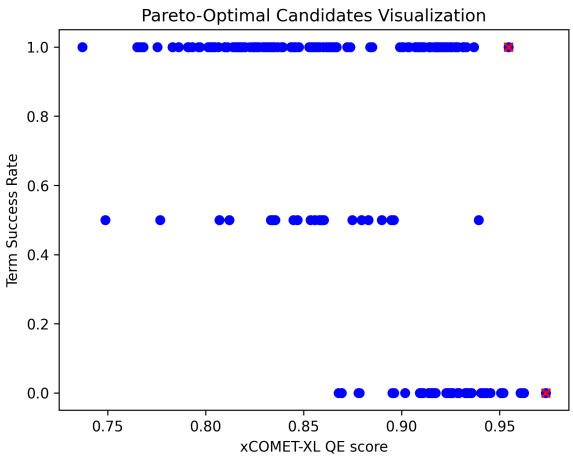

Figure 1: Visualization of Pareto-Optimal Decoding for a sample sentence. Each blue point represents one of 100 translation candidates. The red $\times$ markers highlight the non-dominated solutions. The top marked solution would be chosen by our algorithm to maximize the TSR score.

3.4 Terminology Aware Fine-Tuning

We fine-tuned the `EuroLLM-9B-Instruct` model for English $\rightarrow$ German, English $\rightarrow$ Spanish, and English $\rightarrow$ Russian terminology translation tasks. For each language pair, we used a training set of 200,000 sentence pairs randomly sampled from the OPUS corpora (Tiedemann and Nygaard, 2004).

To prepare data, we adopted a similar approach to the Target Lemma Annotations (TLA) method proposed in (Bergmanis and Pinnis, 2021). Specifically, we selectively sampled nouns and verbs from the target sentences via POS-tagging using `Stanza` (Qi et al., 2020). These words were then aligned with their corresponding source terms using the `fast_align` tool (Dyer et al., 2013), creating parallel term pairs for each sentence. These sentences and term pairs were then formatted using the prompt template shown in Listing 1.

Fine-tuning was performed using LoRA (Hu et al., 2021) over 3 epochs on 4$\times$A100 GPUs conducted through the Oumi (Oumi Community) framework. The specific training hyperparameters are detailed in Table 2.

3.5 Pareto-Optimal Decoding

In the final stage, we used epsilon sampling with $\epsilon = 0.02$ and $T = 1$ to generate 100 candidate translations for each source sentence. This method has previously been found to be effective for creating diverse samples for techniques such as MBR decoding and QE reranking (Freitag et al., 2023). Next, we scored each source-candidate pair using several QE metrics, namely `xCOMET`[2] (Guerreiro et al., 2024), `ReMedy`[3] (Tan and Monz, 2025) and `MetricX`[4] (Juraska et al., 2024). Additionally, we calculated the Term Success Rate (TSR) (Semenov et al., 2023) by verifying the presence of the lemmatized source term words within the lemmatized candidate translation sentence.

Previous research has shown that, while methods such as QE reranking and MBR decoding significantly improve translation quality, they can lead to overfitting to the utility metric (Pombal et al., 2025). We anticipate that simply reranking according to TSR could result in a substantial decrease in translation quality, particularly since TSR is based on matching lemmatized terms and does not consider their grammatical correctness or the quality of the entire translation. Therefore, a more sophisticated selection strategy is required – one that can maximize the task-specific TSR metric while maintaining overall translation quality.

To achieve this balance, we propose an approach based on Pareto optimality, which we name Pareto-Optimal Decoding. This method identifies the set of candidates that represent the optimal trade-off between general quality, as measured by QE scores, and term accuracy, as measured by TSR. Based on the previously calculated metrics, we pruned the translation candidates to a set of Pareto-optimal hypotheses using the `paretoset`[5] library. From this Pareto set of non-dominated solutions, we selected

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

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

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

[5] https://github.com/tommyod/paretoset

System	English → German		English → Spanish		English → Russian	
	COMET	**TSR**	**COMET**	**TSR**	**COMET**	**TSR**
Replace	0.8756	0.7882	0.8819	0.9186	0.8570	0.9168
+ new prompt	0.8725	0.8402	0.8764	0.9302	0.8422	0.9304
+ few shot	**0.8862**	0.7846	0.8935	0.9264	0.8764 *	0.9304
+ LoRA + new prompt	0.8834	**0.8528**	0.8980 *	0.9457	0.8851 *	**0.9497**
+ LoRA + new prompt + few shot	0.8853	0.8474	**0.9016**	**0.9477**	**0.8904**	**0.9497**

(a) Translation quality results per language pair.

System	COMET	BLEU	chrF	TSR
Replace	0.8715	40.70	68.34	0.8745
+ new prompt	0.8637	40.59	68.25	0.9003
+ few shot	0.8854 *	43.91 *	70.38 *	0.8805
+ LoRA + new prompt	0.8888 *	45.53 *	70.80 *	**0.9161**
+ LoRA + new prompt + few shot	**0.8924**	**46.24**	**71.34**	0.9149

(b) Macro average results for all language pairs (English → German, English → Spanish and English → Russian).

Table 3: Ablation tests on translation quality for the WMT25 terminology development dataset, comparing individual and combined gains of each method. Results marked with an asterisk (*) are statistically significant compared to the previous method results.

the candidate that maximizes TSR. This final step ensures that our selection directly addresses the primary objective of the Shared Task, while filtering out suboptimal candidates in terms of translation quality as measured by neural metrics.

Figure 1 illustrates our Pareto-Optimal Decoding method for a single source sentence. For the sake of visual clarity, we limited the method to using only two metrics. Each blue circle corresponds to one of the 100 translation candidates, plotted according to the TSR score on the y-axis and `xCOMET` score on the x-axis. The two red × markers highlight the non-dominated solutions.

Interestingly, the hypothesis yielding the highest `xCOMET` score omits the required terminology entirely, resulting in a TSR score of 0.0. This finding emphasizes the limitations of single-metric optimization and the need for a multi-objective approach for translation quality optimization.

4 Results

Initially, we conducted experiments on the WMT25 development dataset. Table 3 shows the improvements gained by using prompt engineering methods and fine-tuning the model using LoRA. Table 3a summarises the results for each language pair using the `COMET`[6] and TSR metrics. Table 3b shows the macro-averaged values for all language pairs, including the `BLEU` (Papineni et al., 2002) and `chrF` (Popović, 2015) metrics.

The results obtained using the replace method were used as our baseline. As Table 3 shows, using a new prompt slightly decreased the general metrics, but improved the term accuracy metric by an average of approximately 2.5 points. Using a few-shot prompt improved results across all general metrics except TSR. Subsequently, applying LoRA further enhanced translation quality, particularly when the model was used with the new prompt and few-shot examples. Although gains for different methods vary between language pairs depending on dataset characteristics, the averaged values indicate that this combined approach yields the best results.

In addition, 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 $p = 0.05$. Each subsequent processing stage was compared to the previous one. Statistically significant differences are marked with an asterisk (*) in Table 3. The results show that adding few-shot prompting to the baseline solution significantly improved the `COMET` scores for the English → Russian pair, as well as the `COMET`, `BLEU`, and `chrF` scores for the combined dataset for all three language pairs. The second method, which significantly improved the results on `COMET` for the English → Spanish and

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

System		chrF ↑	MetricX ↓	ReMedy ↑	xCOMET ↑	TSR ↑
Replace		**68.34**	2.23	0.6203	0.9243	0.898
Pareto	TSR + xCOMET	64.19	1.77	0.6354	**0.9564**	**0.990**
	+ MetricX	64.18	**1.44**	0.6399	0.9524	**0.990**
	+ ReMedy	66.47	1.62	**0.6567**	0.9540	**0.990**
	+ MetricX + ReMedy	65.69	1.53	0.6487	0.9489	**0.990**

Table 4: Comparison of the use of various metrics in Pareto-Optimal Decoding on the WMT25 terminology development dataset. The results are macro-averaged for each language pair.

Mode	xCOMET-QE ↑	ReMedy-QE ↑	MetricX-QE ↓	TSR ↑
English → German				
noterm	0.9907	0.6481	0.6877	0.2413
proper	0.9770	0.6420	1.1579	0.9903
random	0.9798	0.6458	1.0650	0.9913
English → Spanish				
noterm	0.9803	0.6501	1.6306	0.4015
proper	0.9536	0.6472	1.9855	0.9980
random	0.9586	0.6532	1.9721	0.9980
English → Russian				
noterm	0.9844	0.6290	1.3186	0.3113
proper	0.9567	0.6274	1.7454	1.0000
random	0.9624	0.6318	1.5853	0.9980

Table 5: Evaluation of our final submission to the WMT25 Terminology Translation Task. We calculate the TSR for the noterm mode against terminology constraints in the proper mode as suggested by the task organizers.

English → Russian language pairs, as well as on the entire dataset according to the `COMET`, `BLEU`, and `chrF` metrics, was the method using LoRA with a new prompt. The other methods yield improvements when the averaged metric results are compared, but these are not statistically significant.

Table 4 shows the results of the Pareto-Optimal Decoding experiments. Various metrics were tested to evaluate the candidates in this processing stage. The `xCOMET`, `MetricX` and `ReMedy` metrics were employed and combined with the TSR metric. The results demonstrate that Pareto-Optimal Decoding increases the TSR metric value to 0.99 across all considered translation directions, regardless of the selected metrics. Furthermore, we observe an improvement across all translation quality metrics. However, it's important to note that due to metric interference (MINT) phenomenon (Pombal et al., 2025), the evaluation on utility metrics used during Pareto-Optimal Decoding may yield biased results. We hypothesize that our multi-objective approach mitigates the effects of MINT by encouraging the selection of more robust translation candidates. Based on this analysis, we have decided to use `xCOMET` combined with `ReMedy` in the final solution, leaving `MetricX` for fair evaluation.

Experiments showed that the best results were achieved by using a fine-tuned model together with a modified prompt and few-shot examples, as well as by employing the Pareto-Optimal Decoding method along with the `xCOMET` and `ReMedy` metrics. These methods were used to translate the WMT25 test dataset, except for few-shot prompting, which was found not to affect the translation quality when combined with Pareto-Optimal Decoding. The final results are presented in Table 5. For the `proper` and `random` modes, we utilized our full system. For the `noterm` mode, which lacks a terminology list, we used a baseline model with a modified Pareto-Optimal Decoding that relied solely on QE metrics (`xCOMET` and `ReMedy`).

The final results of all the calculated metrics demonstrate that the proposed method performs well across all evaluated metrics. In particular, TSR achieves values above 0.99 in both the `proper` and `random` dataset modes across all considered directions. In English → Russian direction it even reaches 1.0 in the `proper` dataset mode, which means that the entire specified terminology in the dataset was transferred correctly.

References

Adrien Barbaresi. 2021. Simplemma. Archived snapshot of all versions of Simplemma.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. 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*.

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).

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. 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.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023b. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Tommi Nieminen. 2023. OPUS-CAT terminology systems for the WMT23 terminology shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 912–918, Singapore. Association for Computational Linguistics.

Artur Nowakowski and Krzysztof Jassem. 2021. Neural machine translation with inflected lexicon. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 282–292, Virtual. Association for Machine Translation in the Americas.

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.

Oumi Community. Oumi: an Open, End-to-end Platform for Building Large Foundation Models.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages.

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.

Ryokan Ri, Toshiaki Nakazawa, and Yoshimasa Tsuruoka. 2021. Modeling target-side inflection in placeholder translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 231–242, Virtual. Association for Machine Translation in the Americas.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).