Terminology-Constrained Translation from Monolingual Data using GRPO

Javier Garcia Gilabert[1] **Carlos Escolano**[1,2] **Xixian Liao**[1] **Maite Melero**[1]
[1]Barcelona Supercomputing Center
[2]Universitat Politècnica de Catalunya

Abstract

Terminology consistency is essential for high-quality machine translation, especially in domain-specific and professional contexts, where accurate term translation directly impacts usability. This paper presents the submission from the BSC team to the WMT25 Terminology-Aware Translation Task. We propose the use of GRPO (Group Relative Policy Optimization) to adapt translation models using monolingual data only, without requiring parallel corpora. Our reward function jointly optimizes for terminology adherence and overall translation quality, leveraging quality-estimation metrics. Experimental results demonstrate that our method consistently improves terminology translation across three language directions—English to Spanish, German, and Russian—by up to +0.36 T_P points across all evaluated models.

1 Introduction

In recent years, large language models (LLMs) have emerged as the state of the art across a wide range of natural language processing tasks, including machine translation (MT). Owing to their vast number of parameters, these models possess the capacity to handle multiple languages and adapt across diverse domains. However, their large scale also makes them computationally expensive to fine-tune or adapt to specific settings.

In this work, we address the task of terminology-aware machine translation, aiming to produce domain-specific translations without the need for costly fine-tuning. Rather than adapting the model through retraining, we condition it at inference time by incorporating a bilingual glossary into the translation prompt. This approach enables flexible, on-the-fly domain adaptation, assuming that relevant terminology is available.

Another major challenge in adapting LLMs to MT tasks lies in the scarcity of high-quality parallel corpora, particularly for low-resource language pairs. In many cases, obtaining sufficient parallel data for supervised training is infeasible.

Prompt: Translate the following text from English into Spanish. Please ensure the following terminology is used:

- advertising campaigns → acciones publicitarias
- social platforms → redes sociales

English: "Spotlight's investigation notes that other countries also ran advertising campaigns, including Malta, Greece, Albania, Poland, Armenia and France. Some of these were run by the artists themselves via their own social platforms."

Spanish: La investigación de Spotlight señala que otros países también llevaron a cabo acciones publicitarias –entre ellos, Malta, Grecia, Albania, Polonia, Armenia y Francia–. Algunas de estas campañas fueron organizadas por los propios artistas a través de sus propias redes sociales.

Figure 1: Example of prompt for terminology-aware translations with a glossary. In green, source terms in English. In yellow, target translations in Spanish.

This paper presents the Barcelona Supercomputing Center (BSC) submission to the terminology-aware MT task for the first track: *Sentence/Paragraph-Level Translation*. Our system is built around two key contributions:

- Application of Group Relative Policy Optimization (GRPO) (DeepSeek-AI et al., 2025; Shao et al., 2024): We demonstrate how GRPO can be used to efficiently instruct LLMs in terminology-aware translation. Our experiments show significant performance improvements across multiple language pairs and model architectures.
- Leveraging monolingual data with quality estimation: We show that incorporating quality estimation metrics, such as COMET-KIWI (Rei et al., 2022), allows the model to benefit from

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 1335–1343
November 8–9, 2025.

monolingual data alone, eliminating the need for parallel corpora in supervised training.

2 Related Work

Prior work in terminology-aware machine translation has taken several different approaches. A common strategy involves fine-tuning models with terminology constraints. For instance, Kim et al. (2024) extract terminology from training data to build a glossary and fine-tune the model using inputs augmented with extracted terms. Zheng et al. (2024) propose DragFT, a framework combining dictionary-enhanced prompting, retrieval-augmented few-shot selection, and fine-tuning to improve translation in specialized domains. Another line of work focuses on synthetic data generation and post-editing. Moslem et al. (2023) use LLMs to generate bilingual data containing prespecified terminology, which is then used to fine-tune MT models. They further apply LLM-based post-editing to insert missing terms into system outputs that failed to adhere to terminology constraints. Other methods aim to enforce terminology during decoding. Bogoychev and Chen (2023) explore constrained decoding strategies and LLM-based paraphrasing to increase term fidelity, including the use of negative constraints that penalize incorrect term usage. Reinforcement learning (RL) has also been explored as a way to improve terminology translation. Li et al. (2025) integrate RL with word alignment to define reward signals, enabling models to translate key terminology without explicit term detection at inference time. Our work is most closely aligned with this last line of research. The key distinction is that our approach relies solely on monolingual data and inference-time prompting, and is therefore better suited to low-resource settings where parallel corpora may be scarce or unavailable.

3 Methodology

In this section, we will discuss our proposed method to adapt the LLMs to the task of terminology-aware translation, using monolingual data only.

3.1 Data Preparation

Given that our training data is monolingual, we first create a bilingual glossary containing some of the terms from our source text. To do so, given an English source text, we employ the `spaCy` library (Honnibal et al., 2020) to identify candidate terminology phrases. The extraction heuristic combines three types of linguistic units:

- **Named Entities** Matches proper nouns and numerical expressions (e.g., organizations, locations, dates).
- **Noun Phrases** Captures syntactic chunks that often represent key domain-specific concepts.
- **Adverbial Constructions** Extracts adverbs modifying verbs, adjectives, or other adverbs to capture domain-relevant descriptions.

From these candidates, we select up to five non-overlapping phrases per text. Then, each extracted term is individually translated into the target language using the NLLB 3.3B (Costa-jussà et al., 2022) machine translation model. Once the pairs are generated, all examples are formatted as prompts following the template shown in Appendix Figure 4. Figure 1 shows an example for the English-Spanish pair.

3.2 Group Relative Policy Optimization

In order to adapt the LLMs, we employ Group Relative Policy Optimization (GRPO). This technique allows for efficient training using reinforcement learning without the need for an additional critic model (Schulman et al., 2017; Rafailov et al., 2023). In each training step, for each source sentence q, we sample G candidate translations $\{o_1, o_2, \cdots, o_G\}$ from the current policy model π. Then, we optimize the model parameters maximizing the following objective:

$$\frac{1}{G}\sum_{1}^{G}(\min(\nabla_\pi A_i, \text{clip}(\nabla_\pi, 1-\epsilon, 1+\epsilon)A_i) - \beta\mathcal{D} \quad (1)$$

$$\nabla_\pi = \frac{\pi'(o_i|q)}{\pi_{ref}(o_i|q)} \quad (2)$$

where π' is the adapted model and π_{ref} is the original model used for regularization. The objective function is composed of two terms. The first one computes the average of the losses for all outputs o_i generated from source sentence q. Each output's loss is defined as the minimum of the clipped and unclipped division of the output probabilities of the adapted model by the output probabilities of the original model multiplied by the advantage A_i. Finally, the second term of the loss

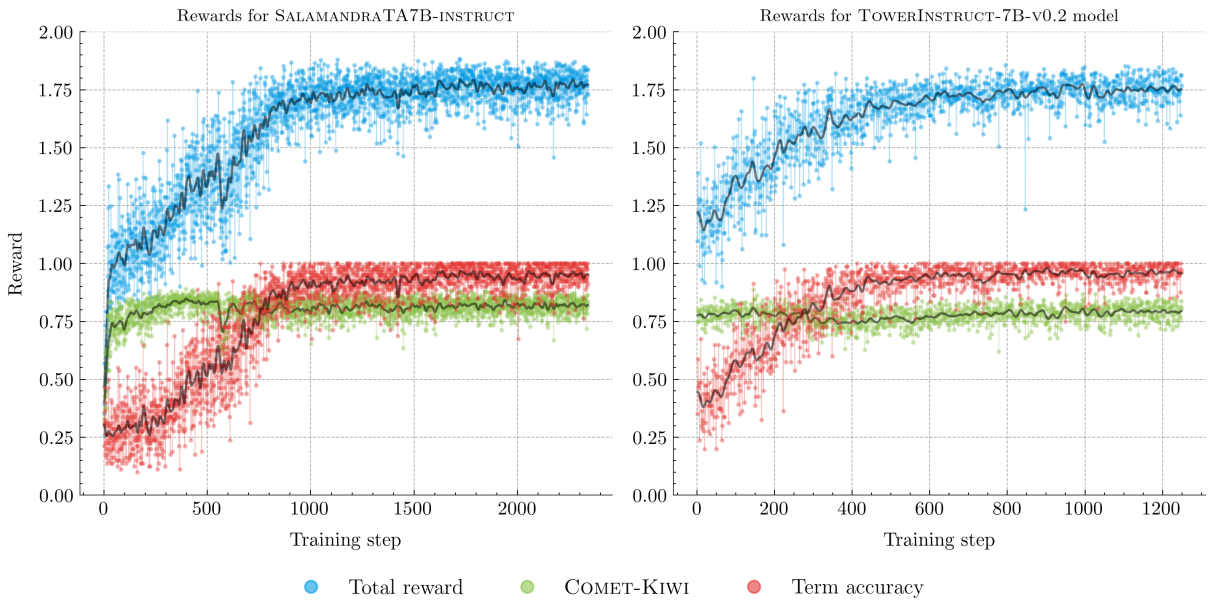

Figure 2: Reward evolution during training for SalamandraTA7B-Instruct (Left) and TowerInstruct-7B-v.02 (Right).

is the Kullback–Leibler distance ($\mathbb{D}_{kl}$) between the output distribution of the adapted model and the original model which is computed as follows:

$$\mathcal{D} = \mathbb{D}_{kl}(\pi' || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi'(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi'(o_i|q)} - 1 \quad (3)$$

Note that there are two hyperparameters that need to be set. First, ϵ controls the PPO clipping threshold. Second, β controls the Kullback–Leibler penalty which measures the relative entropy between both distributions. This penalty prevents the adapted model from diverging too far from the original model, which could cause performance degradation.

The advantage is computed as the normalized reward r_i from each output translation o_i as follows:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}. \quad (4)$$

where r_i denotes the overall reward of o_i. We define our reward function as a sum of two terms: (1) a terminology adherence score that evaluates the correct usage of terms from a provided glossary given in the prompt, and (2) a translation faithfulness score derived from an automated Quality Estimation (QE) metric. This last term is intended to regularize the training process, preventing the model from sacrificing overall translation quality in order to maximize terminology adherence, a behavior that can be seen as reward hacking. The final reward r_i is a linear combination of these two scores:

$$r_i = S_i + \gamma(o_i, q) \quad (5)$$

where S_i denotes the terminology adherence score while $\gamma(o_i, q)$ measures the translation faithfulness of the candidate translation (o_i) with respect to source sentence (q). In this work, we experiment with COMET-Kiwi[1] (Rei et al., 2022) as the quality estimation metric.

Terminology adherence reward The aim of this metric is to compute the proportion of terms in the glossary that are included in the candidate translation. We define the adherence score S_i of a candidate translation o_i as follows:

$$S_i = \frac{1}{|T|} \sum_{i=1}^{|T|} \delta(t_i \in o_i) \quad (6)$$

where T is a glossary of bilingual terms, $|T|$ is the number of terms, $t_i \in T$ is each individual term in the glossary and δ is a transformation of the term to adjust to the translation. For these experiments it will be set to the Identity, but it could be, for example, lemmatization or stemming, allowing the metric to account for morphological variations of the terms (e.g., "run" vs "running"). The score ranges from 0 (no adherence) to 1 (full adherence).

Translation faithfulness reward Despite the adherence score ensuring that the produced translations include the terminology, there are two additional aspects to consider. First, previous work on

[1] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

Direction	Model	T_P	T_F	BLEU	CHRF	COMET
En → Es	TowerInstruct-7B-v0.2	0.48	0.49	27.43	45.83	0.74
	+ GRPO	**0.93**	**0.91**	**51.27**	**74.21**	0.89
	SalamandraTA7B-instruct	0.54	0.54	43.64	62.82	0.79
	+ GRPO	0.90	0.88	47.46	73.75	**0.90**
En → De	TowerInstruct-7B-v0.2	0.60	0.59	38.81	65.43	0.86
	+ GRPO	**0.90**	**0.88**	39.40	68.33	0.87
	SalamandraTA7B-instruct	0.66	0.66	24.57	46.09	0.70
	+ GRPO	0.89	0.87	**44.46**	**71.26**	**0.89**
En → Ru	TowerInstruct-7B-v0.2	0.54	0.57	27.64	58.90	0.87
	+ GRPO	**0.87**	**0.86**	26.08	60.58	0.85
	SalamandraTA7B-instruct	0.66	0.68	20.70	45.10	0.72
	+ GRPO	0.84	0.85	**30.91**	**63.17**	**0.88**

Table 1: Performance of TowerInstruct-7B-v0.2 and SalamandraTA7B-instruct models on terminology-aware translation for English-to-Spanish (En→Es), English-to-German (En→De), and English-to-Russian (En→Ru) directions. Results are reported for both base models and models aligned with GRPO.

using GRPO for Machine Translation (Feng et al., 2025) has observed reward hacking, a phenomenon where models trained on a reward may produce answers that satisfy the reward but fail to solve the tasks. In the terminology task, an example would be copying the glossary without producing a translation. The second concern is that the proposed reward does not take translation quality into consideration, which may lead to catastrophic forgetting of translation quality during training. To prevent these behaviors, we introduce a second term for the reward, where we optimize COMET-KIWI, a quality estimation metric that allows us to evaluate translation quality by computing the similarity between the source sentence and the translation, without requiring a reference translation. As with the terminology adherence score, the faithfulness reward ranges from 0 to 1, with values closer to 1 indicating higher faithfulness.

4 Experiments

4.1 Models

Our experiments required LLMs with strong performance in machine translation. For this reason, we chose two models that were specifically adapted to this task, built on top of generalist LLMs:

TowerInstruct-7B-v.02 (Rei et al., 2024) This model is an adaptation of Llama2-7B for the task of machine translation across ten different languages (English, Portuguese, French, German, Russian, Chinese, Spanish, Dutch, Korean, and Italian). TowerInstruct-7B-v.02 was trained in two main steps: (1) continual pre-training on a combination of monolingual and parallel data; (2) instruction tuning on various tasks such as named entity recognition, machine translation, and post-editing.

SalamandraTA7B-Instruct (Gilabert et al., 2025) This model is an adaptation of Salamandra-7B (Gonzalez-Agirre et al., 2025) for machine translation. It supports 35 languages, including all the official European languages plus several regional Spanish languages such as Catalan, Basque, Galician, and Aranese. The model follows a similar approach to Tower LLM (Alves et al., 2024), with a continual pre-training phase over 424 billion tokens across all supported languages pivoting over Catalan, Spanish, and English. This is followed by an instruction tuning phase on tasks such as paragraph-level translation, post-editing, and alternative translations.

4.2 Evaluation

We evaluate our trained models on two aspects: terminology accuracy and general translation quality. To measure terminology accuracy, we use a provided glossary to compute: (1) Terminology Precision (T_P), the proportion of correctly translated terms, computed using an exact regular-expression match against the reference; and (2) Fuzzy Terminology Precision (T_F), which uses fuzzy matching with an 80% similarity threshold to account

for minor orthographic variations[2]. To assess general translation quality, we evaluate models before and after adaptation using the n-gram-based metrics BLEU[3] (Papineni et al., 2002) and CHRF[4] (Popović, 2015), and the embedding-based metric COMET[5] (Rei et al., 2020) for translation quality.

4.3 Data and Implementation

All experiments were conducted on a subset of the English portion of the *News-Commentary* dataset (Tiedemann, 2012). Model performance was evaluated on the development set released for the WMT25 shared task using the proper terminology subset.

Training was performed with a learning rate of 5×10^{-7} and a temperature parameter of 1.0 for sampling. The maximum generation length was limited to 1024 tokens. To prevent exploding gradients, we applied gradient clipping with a maximum norm of 1.0. Following Feng et al. (2025), we set the GRPO β hyperparameter to 0, and ϵ to 0.3 during training. All models were trained using the `Verl` framework (Sheng et al., 2024) for reinforcement learning, on four NVIDIA H100 GPUs with 64GB RAM each.

4.4 Experimental Results

During training, our first concern was to ensure that the proposed rewards provided enough signal for the model to adapt to the tasks. Figure 2 shows the variation of the two terms of the reward during the training process. We observe that the terminology adherence reward significantly increases during the first training updates, rising from approximately 0.25 accuracy to nearly 1, indicating that, by the end of training, the proposed translations included the terminology in almost all cases.

When looking at the COMET-KIWI score, we observe some differences between the two models. In TowerInstruct-7B-v.02, this reward remains constant throughout training, while SalamandraTA7B-Instruct shows an increase during the first updates of the training. This behavior may be related to greater improvements in translation performance for the latter model. After this initial increase, the COMET-KIWI reward stabilizes.

From these results, we observe that the COMET-KIWI reward functions as a regularization score, maintaining overall translation quality, while the terminology-adherence score is primarily responsible for guiding the model to produce the terminology defined by the glossary.

When looking at the model results in Table 1, we draw similar conclusions. Both models show significant performance gains in terminology accuracy across the directions tested. TowerInstruct-7B-v.02 achieves an average improvement of 0.36 T_P, while SalamandraTA7B-Instruct shows an average improvement of 0.29 T_P.

When looking at the translation performance, we observe more differences between the two models. TowerInstruct-7B-v.02 appears to exhibit inconsistent behavior across languages. While English to Spanish shows significant gains (over 20 BLEU points or 0.15 COMET), English to German shows only small improvements, and in the case of English to Russian even a performance degradation.

Meanwhile, SalamandraTA7B-Instruct shows consistent improvements across all three translation directions. It is worth noting that while the baseline models showed a significant performance gap, this gap narrowed after applying GRPO, with SalamandraTA7B-Instruct even outperforming TowerInstruct-7B-v.02 on both English to German and English to Russian. These differences may be explained by the different behavior of the COMET-KIWI score during training. The improvement observed only in SalamandraTA7B-Instruct may have contributed to its final translation performance.

4.5 Discussion

To test our hypothesis that the translation-faithfulness reward functions as a regularization term, we trained SalamandraTA7B-Instruct using the same training configuration but optimizing only for the terminology adherence reward. Figure 3 illustrates the evolution of this reward in comparison to COMET-KIWI throughout training. While terminology adherence improves at a rate comparable to that observed in Figure 2, translation faithfulness declines rapidly after the initial training steps. This behavior suggests that, in the absence of a faithfulness reward, the model tends to produce translations that diverge semantically from the source sentence, resulting in degraded translation quality and increased hallucinations.

[2]We use the `fuzzywuzzy` Python library for fuzzy string matching.

[3]`Signature: nrefs:1– case:mixed– eff:no– tok:13a– smooth:exp–version:2.3.1`

[4]`Signature: nrefs:1– case:mixed– eff:yes– nc:6– nw:0–version:2.3.1`

[5]https://huggingface.co/Unbabel/wmt22-cometkiwi-da

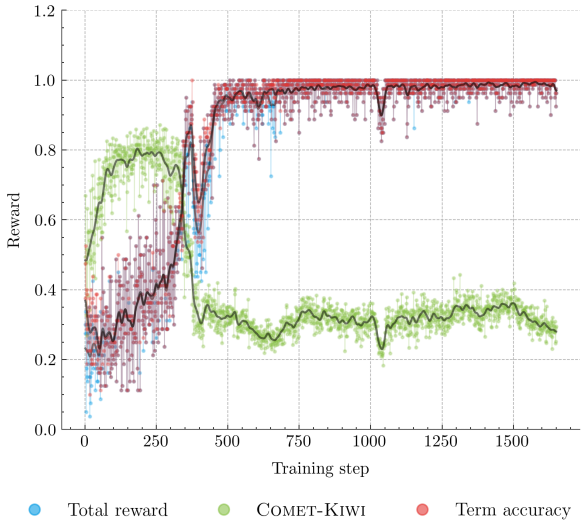

Figure 3: Rewards evolution during training when only the terminology adherence reward is optimized.

5 Conclusions

This paper presents the BSC team's submission to the shared task of terminology-aware Machine Translation. Our results show that GRPO training using only monolingual data can effectively adapt an LLM for this task, producing translations that include the correct terminology in almost all cases. Analysis of the training rewards shows the importance of including a quality-estimation term to regularize training and ensure strong translation performance. Studying the impact of the choice of languages and its relationship with performance remains an avenue for future work.

Limitations

All experiments in this work focus on high-resource languages (English, Spanish, German, and Russian). Quality-estimation metrics can be more consistent for these languages than for other low-resource counterparts. It is worth noting that some extremely low-resource languages may not be supported by any existing quality-estimation model. We leave it for future work to explore the robustness of the model across language families and low-resource scenarios.

Ethical Statement

This work focuses on the term accuracy and overall translation quality of the adapted models. The impact of this adaptation on possible biases, such as gender bias, produced by the system is outside the scope of this study.

All models and datasets used in these experiments are based on publicly available resources, and no direct causes of bias were observed.

Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina Project.

This work has been supported by the Spanish project PID2021-123988OB-C33 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is partially supported by MLLM4TRA (PID2024-158157OB-C32) funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of the ILENIA Project with reference 2022/TL22/00215337.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

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, WMT 2023, Singapore, December 6-7, 2023*, pages 890–896. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. MT-R1-Zero: Advancing LLM-based machine translation via R1-Zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*.

Javier Garcia Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt Argote, Carlos Escolano, and Maite Melero. 2025. From SALAMANDRA to SALAMANDRATA: BSC submission for WMT25 general machine translation shared task. *arXiv preprint arXiv:2508.12774*.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 5 others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642, Miami, Florida, USA. Association for Computational Linguistics.

Zheng Li, Mao Zheng, Mingyang Song, and Wenjie Yang. 2025. TAT-R1: Terminology-aware translation with reinforcement learning and word alignment. *arXiv preprint arXiv:2505.21172*.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. 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.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. Association for Computational Linguistics.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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.

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 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv: 2409.19256*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and*

Evaluation (LREC'12), pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.

A Template

This section presents the template used to prepare instructions for training (Figure 4). We used only one single template. Placeholders:

- { Source Sentence }: source sentence
- { Glossary }: glossary of terms
- { Source Language }: source language name
- { Target Language }: target language name

Template used for training

```
Translate the following text from {Source Language} into {Target Language}.
Please ensure the following terminology is used:
{Glossary}.
{Source Language}: {Source Sentence}
{Target Language}:
```

Figure 4: Example of a template used to construct terminology instructions.