

---

# Breaking Machine Translation with GRPO

---

Aryan Sood<sup>1</sup> Giovanni De Muri<sup>1</sup> Federico Villa<sup>1</sup>

## Abstract

Over the years, machine translation has advanced substantially, to the point that most evaluation examples in existing benchmarks can be translated accurately by state-of-the-art models. As a result, it has become increasingly difficult to identify samples where translation models make significant errors. Since progress in machine translation requires meaningful benchmarks that accurately assess model capabilities and their limitations, automatic techniques are needed to identify and construct difficult test cases. In this work, we introduce MT-GRPO, a reinforcement learning-based approach for training large language models to generate examples that are difficult to translate. Starting from sentences that are easy to translate, our method learns to modify them to increase translation difficulty. Training is performed using GRPO and relies on a combination of reward signals designed to capture key factors that contribute to translation difficulty. Our results demonstrate that MT-GRPO produces high-quality examples that are more challenging for machine translation than the input sentence. Our approach outperforms the base model, demonstrating the effectiveness of the proposed approach.

## 1. Introduction

Current machine translation models achieve very strong performance on standard translation benchmarks, making only a very limited number of errors overall. While this reflects the high quality of these models, it also makes further progress increasingly difficult to measure. Indeed, existing benchmarks have become less informative for evaluating model performance, as most competitive models achieve similarly high scores. Consequently, identifying examples

<sup>1</sup>ETH Zurich, Switzerland. Correspondence to: Giovanni De Muri <gdemuri@student.ethz.ch>, Aryan Sood <arsood@student.ethz.ch>, Federico Villa <fvilla@student.ethz.ch>.

<sup>2</sup>Code available at: <https://github.com/federicovilla55/MT-GRPO>

Submission to the Deep Learning 2025 class., Zurich, Switzerland. 2025. Copyright 2025 by the author(s).

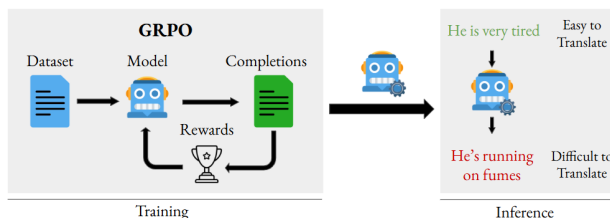


Figure 1. Full pipeline of MT-GRPO. The model is first trained using GRPO with rewards guiding the generation towards high-quality difficult-to-translate sentences. At inference time, given a starting sentence, the model is then able to generate a more complex version of it.

where these models struggle and understanding their weaknesses has become challenging. It is therefore essential to develop new benchmarks that are not easily solvable by state-of-the-art models.

Several approaches have been proposed to construct more challenging test sets. Kocmi et al. (2024) analyzes natural datasets to identify difficult-to-translate examples. However, since most of the natural occurring sentences are relatively simple to translate, this strategy is limited. Alternatively, Pombal et al. (2025) generates difficult to translate examples from scratch in a zero shot manner, using LLMs. However, this approach lacks sufficient difficulty and diversity.

In this work, we address these limitations by training a generator model specifically designed to increase translation difficulty. We propose MT-GRPO, a reinforcement learning-based approach that trains a model to transform easy-to-translate sentences into more challenging ones. Training is performed via GRPO, and it explicitly encourages modifications that increase translation difficulty. By relying on reward driven modifications, rather than unconstrained generation, the modified examples are able to remain semantically close to the originals, and preserve high-quality. Using this approach, we can then systematically modify existing translation benchmarks to increase their difficulty, and produce informative test sets for evaluating translation systems.

## 2. Related Work

**Difficult-to-Translate Examples in Machine Translation**  
To construct benchmarks containing difficult-to-translate

examples, multiple strategies have been proposed. One approach focuses on automatically selecting a subset of examples from existing datasets, based on their translation difficulty (Kocmi et al., 2024). Similarly, Xu et al. (2025) identifies topics that are particularly challenging for translation, and samples sentences related to the most difficult ones from large scale corpora such as the web. However, approaches that rely on naturally occurring data are limited by the low number of challenging translation examples in the wild. A different line of work relies on LLMs to automatically generate difficult translation examples from scratch in a zero-shot manner (Pombal et al., 2025). While this strategy is highly flexible, it often struggles to ensure both sufficient difficulty and diversity in the generated examples. Bridging the two ideas, Zouhar et al. (2025) proposes to start from sentences in existing datasets and iteratively modify them using LLMs, with the goal of increasing translation difficulty at each step. Our work builds on this intuition. However, instead of relying on an iterative procedure, we train a dedicated generator model explicitly optimized to increase the translation difficulty of a given sentence.

### Adversarial Examples with Reinforcement Learning

Reinforcement learning has been used to generate adversarial examples across different settings. In text classification, (Vijayaraghavan & Roy, 2020) uses Self-Critical Sequence Training (SCST) to train a model that introduces minimal perturbations to sentences while preserving their semantic content, in with the goal of fooling a classifier. In the context of machine translation, (Zou et al., 2020) uses an actor-critic reinforcement learning framework to apply token-level edits that degrade translation quality, while preserving the meaning of the original text.

## 3. Method

We now describe our approach for generating text that is difficult to translate, starting from a given sentence.

**Prompting** To guide the model toward generating sentences that are more challenging to translate, we prepend each input text with a prompt. The prompt, provided in Appendix B, instructs the model to modify the input to increase translation difficulty, while preserving the original meaning and length.

**GRPO and Rewards** We train the generator model using GRPO (Shao et al., 2024). Starting from a given dataset, each input is first modified using the prompting strategy described above, after which the model generates candidate completions. Each generated text is then evaluated by a reward model. The reward signal is defined as a weighted combination of multiple components. Together, these rewards encourage increased translation difficulty while pre-

serving semantical content and grammatical correctness. In addition, we include explicit terms to mitigate potential reward hacking, namely length control and penalties for invalid characters. Specifically, we use the following reward components:

- **Translation Difficulty:** To estimate the translation difficulty of the generated text, we typically use a lightweight Sentinel model (Perrella et al., 2024). For some experiments, instead, we employ Comet (Rei et al., 2020), that uses actual translations to assess the difficulty. We used as translation models Helsinki (Tiedemann & Thottingal, 2020), and NLLB (Team et al., 2022). For both Sentinel and Comet, lower values corresponds to higher translation difficulty.
- **Semantic similarity:** To preserve the meaning of the text, we compute the cosine similarity between the embeddings of the original and generated sentence, using an external embedding model (Wang et al., 2020). We penalize generations with low cosine similarity.
- **Grammatical correctness:** To ensure grammatical correctness, we use LanguageTools (lan, 2026) to detect grammar errors and spelling mistakes. Generations are penalized according to the number of grammatical errors in them. Example of generations without this reward are reported in Appendix H.
- **Relative length:** To prevent reward hacking through very short or very long generations, we compare the number of words in the generated sentence with those in the original one. The model is rewarded for producing sentences of similar length.
- **Illegal characters:** To avoid increased translation difficulty through the use of non-standard characters, we penalize generations based on the number of non-ASCII characters in them.

Further details on the GRPO implementation are reported in Appendix A.

## 4. Experiments and Results

We now describe the experimental setup and discuss the obtained results.

**Setup** We focus on the English  $\rightarrow$  Italian language direction. All models are trained on the Tatoeba dataset (Tatoeba Contributors, 2025) for one epoch, using a batch size of 8 and a learning rate of  $1e-5$ . For GRPO training, we sample 8 completions per prompt, using temperature 0.7, top- $p$  0.9, repetition penalty 1.4, and a maximum generation length of 256. Reward weights depend on the translation difficulty estimator used during training. Specifically, we used ( $\alpha_{td} = 2, \alpha_{ss} = 1.5, \alpha_{gc} = 1.5, \alpha_{rl} = 1.5, \alpha_{ic} = 2$ ) for Sentinel and ( $\alpha_{td} = 0.4, \alpha_{ss} = 0.15, \alpha_{gc} = 0.2, \alpha_{rl} = 0.15, \alpha_{ic} = 0.1$ ) for Comet.

**Evaluation** The evaluation set consists of 100 randomly sampled English sentences from WMT19 (Barrault et al., 2019). For each sentence, the model produces 10 candidate modifications, which are evaluated using Sentinel (Perrella et al., 2024), and Comet (Rei et al., 2020), with translation models Helsinki (Tiedemann & Thottingal, 2020), and NLLB (Team et al., 2022). The candidate achieving the highest combined score is selected as the final output. For the selected output, we further evaluate a diverse set metrics capturing various linguistic properties and statistics of the generated sentence. We follow an evaluation protocol similar to that of Zouhar et al. (2025). Specifically, we compute the following metrics: Unique Topics, Cosine Similarity (corpus), Cosine Similarity (original), Grammatical Correctness, Entropy, RIX, Word and Sentence Length, Syntax Complexity, Word Rarity and Naturalness. More details on the evaluation protocol are reported in Appendix C

**Main results** We first evaluate our method by training a Qwen3-4B model using Sentinel as the translation difficulty estimator during GRPO. We compare the results against two baselines: the original sentence left unchanged, and zero-shot generation with the same model without fine-tuning. The results are reported in Table 1. Our method consistently increases the translation difficulty of the original input sentences, as reflected by lower Sentinel and Comet score. This result is also supported by the complexity metrics, which indicate a generally more complex sentence. At the same time, the output quality remains high, as indicated by an high naturalness score and a low number of grammatical errors. The generated sentences are also sufficiently diverse, while remaining close to the original inputs. Compared to the zero-shot baseline, the trained model achieves a higher COMET score while obtaining a substantially lower Sentinel score. This highlights the importance of choosing an appropriate difficulty estimator during training that correctly models translation complexity. Nevertheless, the baseline model generally produces a bit simpler sentences, as indicated by the complexity metrics.

**Training** In Figure 2 and Figure 4, we report the training dynamics of Qwen3 4B trained with GRPO, showing both the total reward and the corresponding loss. The total reward increases steadily over training, with a sharp improvement during the initial training phase, followed by a more gradual trend, indicating that the model quickly aligns with the reward signals, and converges. The behavior of the loss curve mirrors the reward curve: the loss decreases rapidly at the beginning of training, followed by a slower later decline. The individual reward components are reported in Appendix D. We observe consistent improvements across all components, indicating that the optimization is not dominated by a single reward term, but instead that the model improves the objective in a balanced way.

Metric	Q-4B-S	Q-4B-Base	Original
<i>Diversity</i>			
Unique topics	209	166	143
Cosine similarity (corpus)	0.07	0.09	0.05
Cosine similarity (original)	0.79	0.69	1.00
<i>Grammatical Correctness</i>			
Grammatical Errors	0.64	0.53	0.00
<i>Complexity</i>			
Entropy	1.41	0.91	0.65
RIX	19.89	12.79	4.39
Average Word Length	6.78	7.11	5.07
Average Sentence Length	45.42	27.19	18.05
Syntax Complexity	73.60	68.80	31.05
Word Rarity	42.85	48.05	22.70
Naturalness	68.90	67.75	89.15
<i>Translation Difficulty</i>			
Sentinel	-0.71	-0.16	0.13
Comet	0.84	0.83	0.85

Table 1. Comparison of evaluation metrics across multiple models. Specifically, we compare sentences generated by Qwen3 4B trained with our method (Q-4B-S) and by the base Qwen3 4B model (Q-4B-Base) against the starting sentences in the evaluation dataset (Original).

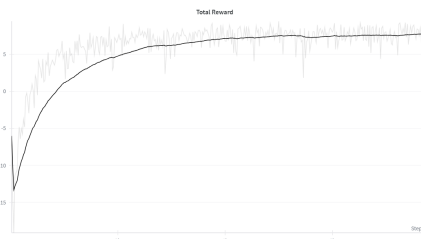


Figure 2. Total reward during GRPO training of Qwen3-4B using Sentinel as the translation difficulty estimator (Q-4B-S).

**Different Models** We evaluate the performance of our method across different model families and model sizes. Specifically, we test our approach using Qwen3 1.7B and Llama3.1 8B Instruct. The results are reported in Table 2. For both models we consistently achieve high translation difficulty scores under both Sentinel and Comet. Surprisingly, the Q-1.7B-S model appears to outperform larger models. However, this apparent advantage does not reflect higher output quality. A qualitative inspection reveals that the model often produces longer sentences that are incoherent and of poor quality, as indicated also by lower naturalness scores and higher entropy values. In contrast, the L-8B-S model shows consistent improvements over the Qwen-4B model. This is due the increased model capacity, that allows for improved sentence generation, while still maintaining high translation difficulty.

**Comet as difficulty estimator** We evaluate the use of COMET as the translation difficulty estimator during GRPO training. Due to computational constraints, we train only the Qwen3-1.7B model under this setting. Its performance is reported in Table 2. Overall, training with Comet yields

Metric	Q-4B-S	Q-1.7B-S	L-8B-S	Q-1.7B-C
<i>Diversity</i>				
Unique topics	209	154	156	162
Cosine similarity (corpus)	0.07	0.23	0.07	0.10
Cosine similarity (original)	0.79	0.58	0.78	0.77
<i>Grammatical Correctness</i>				
Grammatical Errors	0.64	0.53	0.59	0.54
<i>Complexity</i>				
Entropy	1.41	2.65	0.70	1.35
RIX	19.89	19.05	8.85	7.75
Average Word Length	6.78	5.62	6.25	5.96
Average Sentence Length	45.42	61.59	23.24	26.10
Syntax Complexity	73.60	91.95	64.70	55.10
Word Rarity	42.85	77.15	49.75	34.60
Naturalness	68.90	26.75	58.40	79.75
<i>Translation Difficulty</i>				
Sentinel	-0.71	-2.38	-0.37	-0.45
Comet	0.84	0.74	0.78	0.78

Table 2. Comparison of evaluation metrics across multiple models and training rewards. Specifically, we compare generated sentences from Qwen3 4B (Q-4B-S), Qwen3 1.7B (Q-1.7B-S), and LLaMA 3.1-8B (L-8B-S) trained using the Sentinel reward, as well as Qwen3 1.7B trained with the Comet reward (Q-1.7B-C).

Model	Metric	WMT25	WMT19
Original	Sentinel	-1.56	0.13
	Comet	0.65	0.85
Q-4B-S	Sentinel	-2.02	-0.71
	Comet	0.72	0.83

Table 3. Comparison of Sentinel and COMET scores between original sentences and sentences generated by Qwen3-4B GRPO (Q-4B-S), on WMT25 and the simpler WMT19 dataset.

strong results. The model achieves high translation difficulty despite using simpler syntactic constructions, shorter sentences and more common vocabulary. Compared to Sentinel based models, the generated outputs appear more natural, as confirmed by a qualitative inspection. Indeed, Comet mitigates the issues observed for Q-1.7B-S (Section 4). These results show that using Comet leads to better performance at equal model size, highlighting the importance of selecting a good difficulty estimator for training.

**Starting sentence difficulty** We examine how the translation difficulty of the starting input sentences influences the complexity of the outputs. Specifically, we analyze whether simple inputs restrict the model’s ability to generate challenging sentences, and whether already complex inputs can be made even more difficult. We evaluate our model on 100 randomly sampled English sentences from WMT25 (Kocmi et al., 2025), a more difficult benchmark than WMT19, as confirmed by the lower Sentinel and COMET scores reported in Table 3. The results show that starting from more difficult input sentences leads to substantially more challenging outputs. While simpler inputs allow for larger relative improvements, the resulting difficulty remains lower than when starting from already complex sentences.

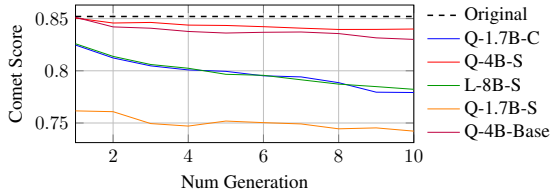


Figure 3. Comet scores as a function of the number of generated sentences per input prompt, across different models.

**Number of iterations** We analyze the effect of the number of sampled generations per input on the translation difficulty estimate. In our evaluation, we generate 10 candidate sentences per input, and select the best one based on the aggregate Sentinel and Comet score. We now study how the score changes as the number of generated samples per prompt is reduced. The results are shown in Figure 3 and Figure 10. Generating more candidate sentences consistently improves performance across all tested models, showing the benefits of diverse generations for producing more complex outputs.

### 5. Conclusion

In this work, we address the challenge of constructing more challenging test sets for machine translation, as current benchmark are no longer sufficiently complex for state-of-the-art models. We propose a GRPO based method to generate difficult-to-translate sentences, starting from input text. Empirical results show that our approach effectively generates more challenging sentences than the baselines. At the same time, the generated outputs are of high quality and largely preserve the original input context. Further evaluation of our method across different model families and sizes demonstrate the robustness and general applicability of our method.

**Limitations** Due to computational constraints, in our analysis we use relatively small LLMs, whose performance is inherently limited. Among the tested models, the larger Llama 8B achieves the strongest results, suggesting that further scaling could yield more challenging generations. Moreover, our method highly depends on the quality of the translation difficulty estimate used during GRPO training. However, better estimates are more computationally demanding.

**Future work** Exploring more advanced translation difficulty estimates for GRPO training, such as MetricX (Juraska et al., 2024), could allow us to obtain more complex generations. Human evaluation would also be valuable to better assess the quality and difficulty of the generated sentences, and identify possible limitations of our method. Finally, extending our approach with an iterative generation loop that progressively increases difficulty, similarly to Zouhar et al. (2025) represents a promising avenue for future work.

## References

- language\_tool\_python: Python grammar checker wrapper for languagetool, 2026. URL [https://pypi.org/project/language\\_tool\\_python/](https://pypi.org/project/language_tool_python/).
- Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. Findings of the 2019 conference on machine translation (WMT19). In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., N ev ol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K. (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL <https://aclanthology.org/W19-5301/>.
- Hansen, L., Olsen, L. R., and Enevoldsen, K. Textdescriptives: A python package for calculating a large variety of metrics from text. *Journal of Open Source Software*, 8(84):5153, April 2023. ISSN 2475-9066. doi: 10.21105/joss.05153. URL <http://dx.doi.org/10.21105/joss.05153>.
- Juraska, J., Deutsch, D., Finkelstein, M., and Freitag, M. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C. (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.wmt-1.35>.
- Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popovi c, M., Shmatova, M., Steingr msson, S., and Zouhar, V. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C. (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.1. URL <https://aclanthology.org/2024.wmt-1.1/>.
- Kocmi, T., Artemova, E., Avramidis, E., Bawden, R., Bojar, O., Dranch, K., Dvorkovich, A., Dukanov, S., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Lakougn, H., Lundin, J., Monz, C., Murray, K., Nagata, M., Perrella, S., Proietti, L., Popel, M., Popovi c, M., Riley, P., Shmatova, M., Steingr msson, S., Yankovskaya, L., and Zouhar, V. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C. (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 355–413, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.22. URL <https://aclanthology.org/2025.wmt-1.22/>.
- Perrella, S., Proietti, L., Scir , A., Barba, E., and Navigli, R. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in!, 2024. URL <https://arxiv.org/abs/2408.13831>.
- Pombal, J., Guerreiro, N. M., Rei, R., and Martins, A. F. T. Zero-shot benchmarking: A framework for flexible and scalable automatic evaluation of language models, 2025. URL <https://arxiv.org/abs/2504.01001>.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL <https://aclanthology.org/2020.emnlp-main.213/>.
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Tatoeba Contributors. Tatoeba: Collection of sentences and translations. <https://tatoeba.org>, 2025. Sentences sourced from the Tatoeba project. Data released under the Creative Commons Attribution 2.0 France (CC-BY 2.0 FR) license.
- Team, N., Costa-juss , M. R., Cross, J.,  elebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzm n, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. No language left behind: Scaling human-centered machine translation, 2022. URL <https://arxiv.org/abs/2207.04672>.

Tiedemann, J. and Thottingal, S. OPUS-MT – building open translation services for the world. In Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., Moorkens, J., Guerberof, A., Nurminen, M., Marg, L., and Forcada, M. L. (eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL <https://aclanthology.org/2020.eamt-1.61/>.

Vijayaraghavan, P. and Roy, D. *Generating Black-Box Adversarial Examples for Text Classifiers Using a Deep Reinforced Model*, pp. 711–726. Springer International Publishing, 2020. ISBN 9783030461478. doi: 10.1007/978-3-030-46147-8\_43. URL [http://dx.doi.org/10.1007/978-3-030-46147-8\\_43](http://dx.doi.org/10.1007/978-3-030-46147-8_43).

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL <https://arxiv.org/abs/2002.10957>. We use the fine-tuned sentence embedding model all-MiniLM-L6-v2.

Xu, W., Zouhar, V., Riley, P., Finkelstein, M., Freitag, M., and Deutsch, D. Searching for difficult-to-translate test examples at scale, 2025. URL <https://arxiv.org/abs/2509.26619>.

Zou, W., Huang, S., Xie, J., Dai, X., and Chen, J. A reinforced generation of adversarial examples for neural machine translation, 2020. URL <https://arxiv.org/abs/1911.03677>.

Zouhar, V., Xu, W., Riley, P., Juraska, J., Finkelstein, M., Freitag, M., and Deutsch, D. Generating difficult-to-translate texts, 2025. URL <https://arxiv.org/abs/2509.26592>.

## A. GRPO

In this section, we explain how the GRPO algorithm works. Consider the current policy  $\pi_\theta$  and the old policy  $\pi_{\theta_{\text{old}}}$ . At each iteration, for each prompt,  $G$  samples are generated using the old policy:

$$\{o\}_1^G = \{o_1, o_2, o_3, \dots, o_G\},$$

where  $\{o_1, o_2, \dots, o_G\}$  are the generated answers, and  $|o_i|$  denotes the length of answer  $o_i$ . These samples are kept fixed for  $L$  iterations, during which the old policy remains fixed and the current policy is updated by optimizing the following objective function:

$$\mathcal{L}(\pi_\theta, \pi_{\text{old}}, \{o_i\}_{i=1}^G) = \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} l_{i,t} - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{old}}).$$

with:

$$l_{i,t} = \min(s_{i,t}(\theta) A_i, \text{clip}(s_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon) A_i),$$

$$s_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})},$$

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

Here,  $r_i$  represents the reward obtained by completion  $o_i$ , computed as a weighted sum of the rewards discussed in Section 3. Intuitively, the model is updated such that a completion that obtained a higher reward, but had a lower probability of generation compared to another completion, becomes more likely to be generated by the updated model. This comparison in the objective function is represented by  $s_{i,t}$ , which is the ratio between the probabilities of a token under the updated policy and the old policy. We calculate the reward for a completion  $o_i$  as follows:

$$r_i = \alpha_1 * r_{td} + \alpha_2 * r_{ss} + \alpha_3 * r_{gc} + \alpha_4 * r_{rl} + \alpha_5 * r_{il}$$

where  $r_{td}, r_{ss}, r_{gc}, r_{rl}, r_{il}$  are the rewards in the same order as defined in Section 3 and  $\alpha_{td}, \alpha_{ss}, \alpha_{gc}, \alpha_{rl}, \alpha_{il}$  the corresponding weights.

## B. Prompts

We instruct the model to increase the translation difficulty of the input sentence using the following prompt:

### Generation Prompt

Please rewrite the following sentence so that it becomes slightly more complex and harder to understand, while still remaining logical and roughly the same length. Do not change the meaning. Only

output the rewritten sentence, no comments or additional text. Only increase the linguistic complexity. Sentence you have to change: SENTENCE

### C. Evaluation Details

We provide further evaluation details, specifying how each metrics computed. The metrics reported are the following:

- Unique topics: Topics are extracted per sentence using an LLM-as-judge, and we report the number of distinct topics across the evaluation set.
- Cosine similarity: Mean pairwise cosine similarity between sentence embeddings, computed using Wang et al. (2020). We compute it both between the entire evaluation set (corpus) and between the generated and original sentence (original).
- Grammatical correctness: Average number of grammatical errors per sentence, measured with Language-Tools (Ian, 2026) .
- Entropy: Mean entropy of the sentence, computed with (Hansen et al., 2023).
- RIX: Mean RIX readability score, computed with (Hansen et al., 2023).
- Word and Sentence Length: Average word length (excluding stopwords) and average sentence length.
- Syntax Complexity: Sentence-level syntactic complexity score obtained using an LLM-as-judge.
- Word Rarity: Lexical rarity score obtained using an LLM-as-judge.
- Naturalness: LLM-as-judge score measuring the likelihood that the sentence is human-written.

Metric scores for unique topics, naturalness, word rarity, and syntactic complexity are obtained using an LLM-as-judge. Specifically, we employ a Qwen3 4B model as the judge, prompted with the prompt below and temperature set to 0.

#### LLM-as-Judge Prompt

Analyze the following text and return the answer in JSON. We want to determine the following attributes: - naturalness: on a scale from 0 (wholly unnatural) to 100 (fully human-like and would occur in a corpus). - word rarity: on average, how rare are the words from 0 (average modern human would use this word daily) to 100 (average modern

human would not understand the word). - syntax complexity: on a scale from 0 (simplest possible sentence) to 100 (most complex and hard to understand). - topics: list of 1 to 5 topics that the text is about. Provide only the output in JSON and nothing else. The output should look like this (no extra backticks or newlines): "naturalness": 80, "word rarity": 50, "syntax complexity": 70, "topics": ["science", "technology"] The sentence to analyze is: SOURCE\_TEXT

### D. Training Curves

In Figure 4, we show the loss dynamic of the Qwen3 4B model trained using Sentinel (Q-4B-S). In Figure 5, Figure 6, Figure 8, Figure 7, and Figure 9, instead, we provide additional figures of each singular reward signal obtained during the GRPO training of the same Q-4B-S model.

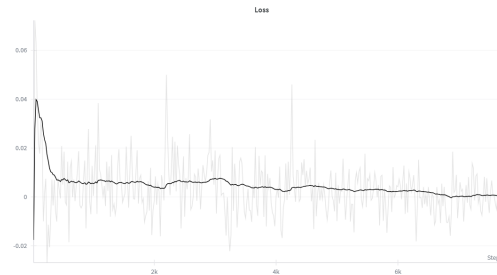


Figure 4. Loss during GRPO training (as defined in Appendix A) of Qwen3-4B using Sentinel as the translation difficulty estimator (Q-4B-S).

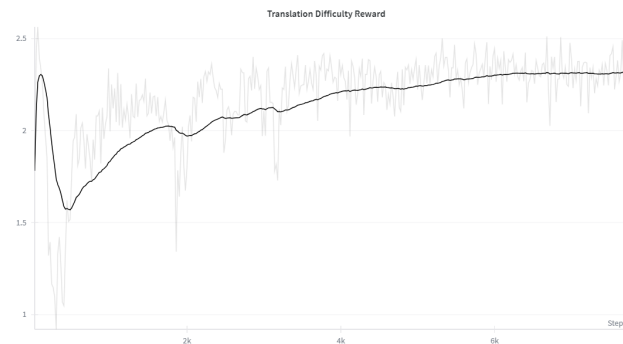


Figure 5. Translation Difficulty reward signal, during GRPO training of the Qwen3-4B model (Q-4B-S).

## Breaking Machine Translation with GRPO

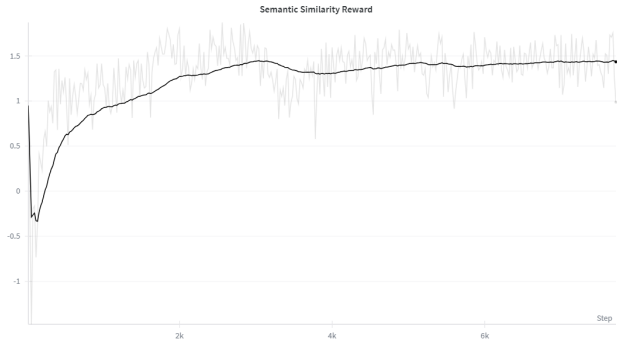


Figure 6. Semantic Similarity reward signal, during GRPO training of the Qwen3-4B model (Q-4B-S).

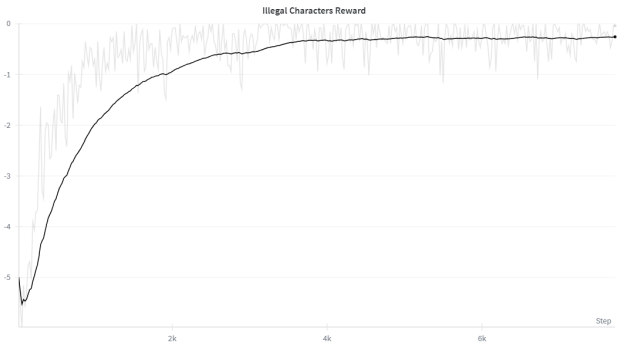


Figure 9. Illegal Characters reward signal, during GRPO training of the Qwen3-4B model (Q-4B-S).

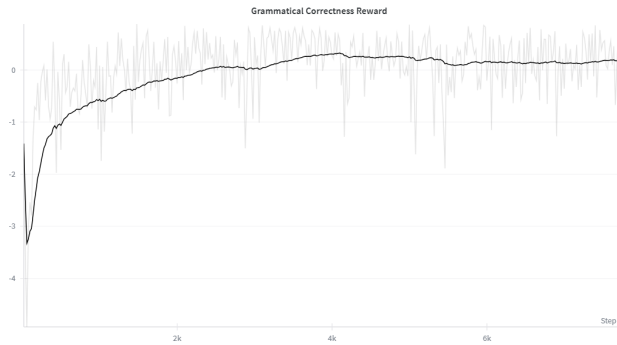


Figure 7. Grammatical Correctness reward signal, during GRPO training of the Qwen3-4B model (Q-4B-S).

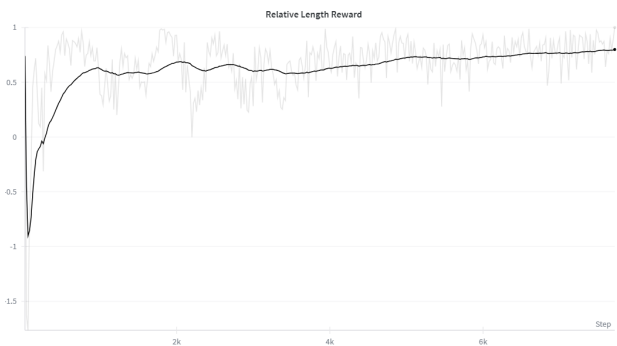


Figure 8. Relative Length reward signal, during GRPO training of the Qwen3-4B model (Q-4B-S).

### E. Evaluation on WMT25

In Table 4, we present the evaluation of the trained models on the WMT25 dataset, a more challenging benchmark than WMT19.

Metric	Q-4B-S	Q-1.7B-S	L-8B-S	Q-1.7B-C	Q-4B-Base	Original
<i>Diversity</i>						
Unique topics	233	210	193	185	235	170
Cosine similarity (corpus)	0.13	0.19	0.15	0.12	0.19	0.11
Cosine similarity (original)	0.84	0.72	0.79	0.88	0.74	1.00
<i>Grammatical Correctness</i>						
Grammatical Errors	2.01	1.39	1.06	2.58	0.86	0.00
<i>Complexity</i>						
Entropy	3.16	5.32	2.94	4.04	3.93	3.66
RIX	28.83	28.61	35.86	5.01	29.62	3.65
Average Word Length	5.93	5.14	6.14	5.11	6.46	4.73
Average Sentence Length	74.51	111.90	104.70	24.18	70.13	21.40
Syntax Complexity	80.71	90.75	87.75	63.85	81.06	49.65
Word Rarity	58.15	73.75	69.75	45.90	57.40	36.15
Naturalness	53.90	32.25	37.45	66.85	64.15	77.25
<i>Translation Difficulty</i>						
Sentinel	-2.02	-2.54	-2.16	-1.82	-1.78	-1.56
Comet	0.73	0.67	0.66	0.66	0.67	0.65

Table 4. Comparison of evaluation metrics across all trained models on the WMT25 dataset.

### F. Multiple iterations

In Figure 10 we report the Sentinel score across different models, as a function of the number of generated sentences per input prompt at inference time.

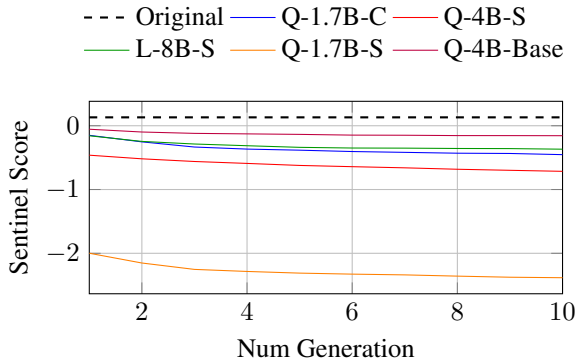


Figure 10. Sentinel scores as a function of the number of generated sentences per input prompt, across different models.

### G. Generation Examples

We present multiple examples of sentences generated by the different models evaluated in this work, namely Q-4B-S, Q-1.7B-S, Q-1.7B-C, and L-8B-S, with inputs taken from either the WMT25 or WMT19 dataset.

#### WMT25 — Example 1

**Original:** *Meghan’s rebranded herself - and it’s another valiant attempt to flog poshness to the little people*

**Q-4B-S:** Meghan’s rebranded herself as a strategic move to promote an image of luxury and refined appeal, which involves targeting the general public with efforts to sell an air of elegance, maintaining a consistent push to flog poshness to the little people through deliberate and ambitious marketing initiatives

**Q-1.7B-S:** Complicatedly, Meghan’s rebranded herself with a challenging effort, yet maintaining the brave idea of selling elegant charm to the common folks, while making it more confusingly intricate, as the effort remains true to the original thought, and the structure is still keeping the same length, without altering the main sense.

**Q-1.7B-C:** Meghan’s rebranded herself with a more determined effort to promote the idea of luxury among those who are not as prominent, and it’s another valiant attempt to flog poshness to the little people.

**L-8B-S:** Meghan rebranded herself alongside it - as well as another valiant endeavor to peddle opulence among the diminutive crowd’s masses

**Q-4B Base:** Meghan has undergone a strategic reinvention—representing yet another concerted effort to project an air of refined elitism toward the more modestly situated.

#### WMT25 — Example 2

**Original:** *Pig Heart Boy review - lively staging of Malorie Blackman’s stimulating novel*

**Q-4B-S:** Review of Pig Heart Boy featuring a vibrant production based on Malorie Blackman’s novel that offers energetic performances and dynamic presentation, highlighting the novel’s engaging content through an active and lively staging approach maintaining the original creative inspiration and enthusiastic execution.

**Q-1.7B-S:** Complicated portrayal of Pig Heart Boy review, intricate presentation of lively staging Malorie Blackman’s engaging book, fascinating performance amid stimulating novel, challenging to grasp yet maintaining the exciting event, remarkable effort in the hard-to-read sentence, vivid depiction of the original thought, elaborate structure keeping the same sense, brilliant concept behind the complex thought, mesmerizing display of the remarkable work, bewildering manner where the logical message remains undisturbed.

**Q-1.7B-C:** The Pig Heart Boy review was given a lively staging of Malorie Blackman’s stimulating novel, which was a bit more engaging than previously thought.

**L-8B-S:** A pig-hearted boy reviews the lively staging of Malorie Blackman’s novel, whose stimulating pages

**Q-4B Base:** A vibrant and dynamic production of Malorie Blackman’s intellectually invigorating novel, the Pig Heart Boy review presents a theatrically charged interpretation.

**Q-1.7B-C:** Musk’s Threats Suddenly Darken as Trump Legal Losses Trigger MAGA Fury are now more intense, with the situation evolving in a way that adds a new layer of concern. The events have sparked a reaction that is both surprising and unexpected, maintaining the original intent while adding a subtle complexity.

**L-8B-S:** Musk’s threats unexpectedly darken alongside Trump’s legal setbacks, prompting the latter’s MAGA fervor as a direct consequence of his electoral defeats

**Q-4B Base:** As Trump’s legal setbacks ignite a surge of MAGA-driven outrage, Musk’s previously tempered pronouncements have undergone a notable intensification, veering into more pronounced and assertive tones.

**WMT25 — Example 3**

**Original:** *Musk’s Threats Suddenly Darken as Trump Legal Losses Trigger MAGA Fury*

**Q-4B-S:** As Trump faces legal setbacks, heightened MAGA enthusiasm intensifies, Musk’s previously cautious statements shift abruptly to more urgent tones, the sudden escalation in threats is directly influenced by Trump’s legal losses, resulting in a broader intensification of political reactions and amplified public concern surrounding Musk’s evolving rhetoric and associated political dynamics.

**Q-1.7B-S:** Complicated Allegories Spark Trump’s Unseen Consequences, Where Musk’s Bold Claims Get Confused, Meanwhile the Nationalist Feelings Rage On, Since the Legal Battles Lose Their Clear Picture, While the Crowd’s Shocking Mistakes Stay Unseen, But the Situation Remains Uncomplicated, Despite the Hardened Understanding of the Public’s Displeasure.

**WMT19 — Example 1**

**Original:** *There is a wonderful view of the sea from the terrace.*

**Q-4B-S:** From the terrace, one can observe a magnificent sea landscape that provides a wonderful outlook, where the view of the sea is prominently displayed and appreciated, maintaining the original delightful characteristics of the terrace’s sea-facing perspective.

**Q-1.7B-S:**asionally, the impressive panorama of the ocean is revealed with a remarkable sight, as the elevated place offers a fascinating landscape, while the pleasant scenery remains clear, but the breathtaking scene is maintained, making it hard to figure out the natural beauty, yet the delightful view stays in place, despite the complicated structure, keeping the same length in a confusing manner.

**Q-1.7B-C:** There is a wonderful view of the sea from the terrace, offering a glimpse that is both inviting and serene. The perspective is framed by a setting that invites reflection, yet the experience remains a simple and elegant sight.

**L-8B-S:** There exists a resplendent view of the sea that unfolds from the elevated platform.

**Q-4B Base:** From the elevated terrace, one is afforded a remarkably expansive and picturesque vantage point overlooking the maritime expanse.

**L-8B-S:** The state relies on its citizens to mount a defense in support of the country

**Q-4B Base:** The jurisdictional administration is presently relying upon its populace to uphold the national integrity through civic vigilance and active participation in safeguarding territorial sovereignty.

#### WMT19 — Example 2

**Original:** *The state is counting on its citizens to defend the country.*

**Q-4B-S:** The state relies upon its citizens' willingness and ability to protect the nation, ensuring national security through active participation in defense efforts, maintaining trust in governmental responsibilities, and depending on public commitment to uphold the country's safety and stability.

**Q-1.7B-S:** Distinctly, the political area relies upon the people's protection, maintaining the military strength, while the nation's shield is being supported, yet the community is expecting to guard the land, making it difficult to grasp, but still keeping the same idea, as the world observes the situation, with the defense effort being more intricate, yet the truth remains clear.

**Q-1.7B-C:** The state's reliance is placed upon its citizens to protect the country with a sense of duty. This acknowledgment of the citizens' role is a testament to their commitment to the nation's security.

#### WMT19 — Example 3

**Original:** *The new age brought new instructions.*

**Q-4B-S:** New instructions introduced during the transition into the new age resulted in changes aligned with the era's developments, reflecting the shifts inherent in the updated time period and the adoption of forward-looking directives within the new age context.

**Q-1.7B-S:** painstakingly intricate era introduced elaborate guidelines with a confusing structure, while maintaining the youthful trend, yet subtly altering the knowledge, and gradually making the rules more unclear, but still keeping the youthful idea in place, making it difficult to follow, yet preserving the original message in a tangled manner.

**Q-1.7B-C:** The evolving era introduced fresh guidelines with a renewed focus on direction. The ongoing shift introduced new rules with a renewed focus on direction.

**L-8B-S:** The introduction of a new era ushered in additional guidelines alongside its own.

**Q-4B Base:** The epochal shift introduced a reconfiguration of procedural directives.

#### WMT19 — Example 4

**Original:** *We have fully grown trees on the land, so we can stay underneath them.*

**Q-4B-S:** The land currently supports mature trees that have reached full development, allowing us to remain sheltered beneath their established canopies, which directly enables continued presence under the fully grown trees we have on the land, ensuring our ability to stay underneath them.

**Q-1.7B-S:** Complicated forest trees have remarkably expanded tall structures on the area, making it possible to keep shelter above, while the natural growth is ensured, but the situation remains clear, making the process difficult to grasp, yet the trees still support the ground in a complicated way, keeping the place safe, and the logic remains untouched, though the sentence becomes more tangled.

**Q-1.7B-C:** The development of trees on the land has been completed, allowing us to remain beneath them in a more intricate manner.

**L-8B-S:** We possess fully mature trees across the terrain, since we are able to remain sheltered beneath their sprawling canopies.

**Q-4B Base:** On the designated territory, where mature specimens of the flora have attained their full developmental stage, we are able to maintain sheltered occupancy beneath their expansive canopies.

thing. However, in French, the word *demande* is a better way “to ask” for something. The speaker is showing more consideration of the other person’s feelings.

**Generated Text:** In Anglais, the action “requisit” is any thing thou exults on account of a robust desire for a thing thine must possess. It’s generally a disquieted manner to utter a thing for this things thrice. Yet, in French, the term *recher* is a pricier approach ‘t o quest’ for este mething. The speaker is expostulating much greater care of the althern person’s sensibilities.

**Tatoeba — Example 2**

**Original:** *We were at breakfast when the Colonel’s butler rushed in with all his propriety shaken out of him. “Have you heard the news, sir?” he gasped. “At the Cunningham’s sir!” “Burglary!” cried the Colonel, with his coffee-cup in mid-air. “Murder!”*

**Generated Text:** We was thar on brekfast wene the Cohn’l’d bute l’chur nith a hte rul’th shtehd oot fum ehm d’him. ”’Hav ye hearth de nees, sair?’ he gaspeth. ‘A tae m’Cunnin’ms siro!’ ‘Brigandage!’ cryed te Cohn’L’S cufep-ke pind i maid-aire. ‘Murrdey!’.

**H. Training without Grammatical Correctness Reward**

We present examples of generations produced by the Qwen3-1.7B model trained with Comet as the difficulty estimator (Q-1.7B-C), when we remove the grammatical correctness reward. The resulting outputs are of low quality, as they are largely non-sensical and exhibit a high number of grammatical errors.

**Tatoeba — Example 1**

**Original:** *In English, the verb “demand” is something you say when you have a strong need for something. It is usually an impolite way to ask for some-*

**Tatoeba — Example 3**

**Original:** *There are also calls to build seawalls to protect the city, which already suffers from frequent flooding. Such a project would cost billions of dollars. And even if the walls are built, nature could still eventually reclaim everything people have built in some unprotected low-lying areas.*

**Generated Text:** There is also use to construct marshals to guard the town, thar be yet often flowin' wetlands for safegrounnd the metropolis, whic h yaerly encom s fréquent submersion, sinn er e d t o n l y r é p u b l i c m o n g f r e q u e n t l o o d s o f t r i l l i o n s o f d o l l o s i r s . A n s u c h a n e f f o r t w o u l d e c o s t b i l l i o n z o f t e r r e n s o f d o l l a r s . A n d e v e n s i f t h e m u r k l e s a r e b u i l t , n a a t u r e c o u l d s t i l l e v e n t u l y r e c o u m e a l l t h i n g s p e o p l e d h a v e b u d d i n s o m e s u n p r o t e c t e d l o w - l i e i n g z o n e s , i n t h e r e x p e n s i o n o f u n k i n d f l u d g i n s o n s o m e i n f u n d e d s h a l - l o w a r e a s c a t e r e .