Improving Low-Resource Machine Translation via Round-Trip Reinforcement Learning

Anonymous ACL submission

Abstract

Low-resource machine translation (MT) has gained increasing attention as parallel data from low-resource language communities is collected, but many potential methods for improving low-resource MT remain unexplored. We investigate a self-supervised reinforcement-learning-based fine-tuning for translation in low-resource settings using round-trip bootstrapping with the No Language Left Behind (NLLB) family of models. Our approach translates English into a target low-resource language and then back into English, using a combination of chrF++ and BLEU as the reward function on the reconstructed English sentences. Using the NLLB-MD dataset, we evaluate both the 600M and 1.3B parameter NLLB models and observe consistent improvements for the following languages: Central Aymara, Friulian, Wolof and Russian. Qualitative inspection of translation outputs indicates increased fluency and semantic fidelity. We argue that our method can further benefit from scale, enabling models to increasingly leverage their pretrained knowledge and continue self-improving.

1 Introduction

Low-resource machine translation remains one of the most persistent challenges in natural language processing. Modern neural translation systems achieve impressive performance for high-resource languages (Vaswani et al., 2017; Conneau et al., 2020; Hendy et al., 2023), where millions of parallel sentence pairs are available for supervised training. However, thousands of languages lack parallel corpora of sufficient scale (Koehn and Knowles, 2017; Haddow et al., 2022), leading to models that perform poorly or fail to generalize outside narrow training domains. As a result, speakers of low-resource languages continue to face limited access to digital information, reduced participation in multilingual technologies, and diminished inclusion in global communication systems.

To address these limitations, we introduce a framework to improve MT system without any parallel corpus that integrates reinforcement learning (RL) via round-trip bootstrapping. Our system operates in a self-supervised manner by translating from English into a low-resource language and then back to English. In this setup, the intermediate low-resource translation is treated as a latent step required to reconstruct the original English sentence. We use RL to optimize this process, rewarding the model based on lexical and semantic metrics such as BLEU and chrF++. This allows the model to improve translation quality in the low-resource language without requiring any human-annotated parallel data.

The necessity for this approach arises from the inherent flaws in standard machine translation training. Most systems rely on maximum likelihood estimation (MLE) (Fisher, 1992), which teaches models to predict the next token given ground-truth history. While effective, MLE suffers from *exposure bias* (Bengio et al., 2015; Wang and Sennrich, 2020; He et al., 2024), arising from discrepancies between training and inference conditions, and an *objective mismatch* (Maksai and Fua, 2018), where the training loss does not directly optimize evaluation metrics like chrF++ (Popović, 2015). In low-resource settings, these issues are amplified because models have fewer "gold" examples to prevent them from drifting during inference.

Furthermore, while back-translation (Sennrich et al., 2016) is currently the dominant method for low-resource MT, its effectiveness is fundamentally constrained by the quality of the reverse model. When initial translations are weak, the synthetic data produced by back-translation can reinforce existing model errors. This creates a feedback loop that limits gains, particularly for morphologically rich or typologically distant languages (McNamee and Duh, 2023). Standard self-supervised approaches lack a mechanism to prioritize high-

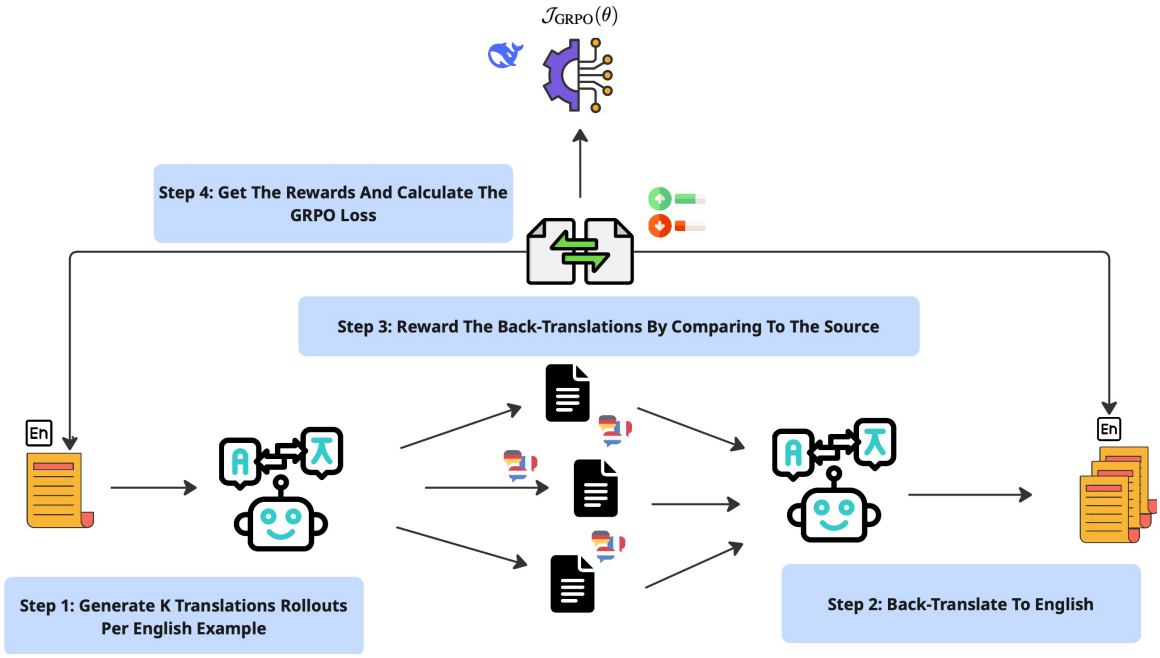

Figure 1: Overview Of Our Low-Resource Machine Translation With Round-Trip Reinforcement Learning Method.

quality synthetic examples or directly optimize task-specific translation metrics.

RL (Sutton and Barto, 2018) provides a natural solution to these bottlenecks. By treating translation as a sequential decision-making process, RL allows models to directly optimize non-differentiable metrics and exposes them to their own predictions during training. While foundational work introduced policy gradient methods (Ranzato et al., 2016; Wu et al., 2018) to reduce exposure bias, and recent studies have applied RL to large-scale models (Feng et al., 2025), the application of RL to bridge the specific gap in round-trip low-resource translation remains relatively under-explored.

In this work, we demonstrate that our RL-based bootstrapping method significantly improves both the adequacy and fluency of translations. Our main contributions are:

- We propose a novel RL-based framework for round-trip translation that operates entirely without parallel data.
- We show that optimizing for reconstruction rewards significantly improves surface-level well-formedness and semantic adequacy in low-resource settings.
- We provide an extensive analysis across several languages.

2 Related Work

2.1 Improving NMT with Monolingual Data and Roundtrip Translations

Back-translation has been a foundational technique for leveraging monolingual data in neural machine translation. Sennrich et al. (2016) introduced back-translation and demonstrated that target-side monolingual data can substantially improve translation quality by generating synthetic parallel data; mixing these synthetic examples with genuine bitext yields strong performance gains and supports effective domain adaptation. Lample et al. (2018) showed that unsupervised machine translation is feasible by combining denoising autoencoding, back-translation, and a cycle-consistency objective that enforces accurate reconstruction of source sentences after a forward–backward translation loop. This line of work treats forward and backward translation models as cooperative components whose outputs must remain mutually consistent. Huang (1990) is an early work that explicitly uses round-trip translation as a quality check for "target language inexpert" users: texts are translated into a foreign language and immediately back so users can spot major errors without knowing the target language. Federmann et al. (2019) extend this idea to paraphrasing, systematically studying "multilingual whispers," where sentences are translated through one or more pivot languages and back to

generate diverse paraphrases, showing that round-trip translation can serve not only for quality control but also as an effective paraphrase generation method.

2.2 Reinforcement Learning and GRPO for NMT

Wu et al. (2018) provided a practical recipe for RL in NMT, showing that multinomial sampling, reward shaping, and objective mixing with MLE lead to consistent gains. Feng et al. (2025) proposed MT-R1-Zero, applying R1-Zero–style RL to MT without supervised warm-starts. Their rule-metric mixed reward improved lexical and semantic quality via GRPO (Shao et al., 2024), suggesting that pure RL can rival SFT for LLM-based translation.

GRPO extends Proximal Policy Optimization (PPO) (Schulman et al., 2017) and removes the value critic.

3 Method

We propose a reinforcement learning framework that leverages round-trip translation consistency to fine-tune pretrained multilingual language models for low-resource machine translation. Our approach combines Group Relative Policy Optimization (GRPO) with back-translation-based rewards, enabling effective adaptation without parallel data augmentation.

Preliminary For each input x, GRPO samples a group of G candidate outputs $\{y_1, \ldots, y_G\} \sim \pi_{\theta_{\text{old}}}(y \mid x)$ from the current policy. The update maximizes

$$\mathcal{J}_{\text{GRPO}}(\theta) = E_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\,\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\Big) - \beta\, D_{\text{KL}}\big[\pi_\theta \| \pi_{\text{ref}}\big]\Bigg],$$

$$\text{where } r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{1}$$

GRPO differs from PPO in how it defines the advantages A_i. Instead of subtracting a critic prediction, GRPO normalizes rewards within the sampled group. Let $R_i = R(x, y_i)$; then

$$A_i = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\})}. \tag{2}$$

When all or none of the candidates solve the problem, every A_i is zero and the policy gradient vanishes except for the KL term.

3.1 Problem Formulation

We consider the setting where *no parallel corpus is available* for the source and target language pair. Instead, we assume access to a pretrained multilingual neural machine translation (NMT) model π_θ and monolingual corpora in the source language (English). Our goal is to improve the translation quality in both directions for this low-resource pair *without* relying on any parallel data for training.

We formulate this adaptation problem as a reinforcement learning problem. At each step, the model samples translation candidates and receives rewards that reflect translation quality, defined purely from intrinsic signals. Concretely, starting from a source sentence x, the model generates a target hypothesis $\tilde{y} \sim \pi_\theta(\cdot \mid x)$, then translates back $\hat{x} \sim \pi_\theta(\cdot \mid \tilde{y})$. The agent is rewarded according to the quality of the round-trip reconstruction, for example via a similarity measure (chrF++) between x and $\hat{x}$. In this way, the model learns from self-consistency and monolingual structure, without any parallel supervision.

Let x denote a source sentence and y denote its target translation. The policy $\pi_\theta(y|x)$ produces candidate translations by autoregressively sampling tokens $y = (y_1, y_2, \ldots, y_T)$ conditioned on the source.

3.2 Round-Trip Translation Reinforcement Training

Our key insight is that translation consistency in both directions provides a stronger learning signal than unidirectional training alone. For each source sentence x, we execute a two-phase generation process:

Phase 1: Forward Translation. Generate K candidate translations from source to target:

$$\hat{y}^{(k)} \sim \pi_\theta(\cdot \mid x), \quad k \in \{1, \ldots, K\} \tag{3}$$

Phase 2: Back-Translation. For each forward translation $\hat{y}^{(k)}$, generate K back-translations to the source language:

$$\hat{x}^{(k,j)} \sim \pi_\theta(\cdot \mid \hat{y}^{(k)}), \quad j \in \{1, \ldots, K\} \tag{4}$$

The model is trained on the back-translation phase, where rewards measure how well the back-translated sentence $\hat{x}$ matches the original source x. This self-consistency objective encourages the model to produce translations that preserve semantic content bidirectionally.

3.3 Reward Design

We employ a reward function combining character-level and word-level translation quality metrics. For a generated back-translation $\hat{x}$ given the original source x:

$$R(\hat{x}, x) = \lambda_{\text{chrF}} \cdot \text{chrF++}(\hat{x}, x) + \lambda_{\text{BLEU}} \cdot \text{BLEU}(\hat{x}, x) \tag{5}$$

chrF++. (Popović, 2015) is a character n-gram F-score that is particularly suitable for morphologically rich and low-resource languages, as it provides smoother gradients than word-based metrics:

$$\text{chrF++} = (1 + \beta^2) \cdot \frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}} \tag{6}$$

BLEU. (Nießen et al., 2000) with effective order provides complementary word-level precision signals, helping the model learn appropriate lexical choices. Algorithm 1 summarizes our training procedure.

Algorithm 1 Round-Trip Translation Reinforcement Training with GRPO

Require: Pretrained model π_θ, source-language corpus $\mathcal{D}$, learning rate η, group size K, KL coefficient β, clip parameter ϵ_c

Initialize reference policy $\pi_{\text{ref}} \leftarrow \pi_\theta$
for epoch $= 1$ to E **do**
  for batch (x) in $\mathcal{D}_{\text{source lang}}$ **do**
    // Forward translation
    $\{\hat{y}^{(k)}\}_{k=1}^{K} \sim \pi_\theta(\cdot \mid x)$ *with temperature sampling*
    // Back-translation
    $\{\hat{x}^{(k)}\}_{k=1}^{K} \sim \pi_\theta(\cdot \mid \hat{y}^{(k)})$
    // Compute rewards
    $R^{(k)} \leftarrow \lambda_{\text{chrF}} \cdot \text{chrF++}(\hat{x}^{(k)}, x) + \lambda_{\text{BLEU}} \cdot \text{BLEU}(\hat{x}^{(k)}, x)$
    // Update policy
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{GRPO}}(\theta)$
    if step $\bmod N_{\text{update}} = 0$ **then**
      $\pi_{\text{ref}} \leftarrow \pi_\theta$
    end if
  end for
end for
return Fine-tuned policy π_θ

4 Experiments

We base our experiments on NLLB-200 (No Language Left Behind) (Goyal et al., 2022), a massively multilingual translation model that covers 200 languages. In particular, we use the distilled 600M and 1.3B parameter variants. NLLB is an encoder-decoder Transformer (Vaswani et al., 2017) with language-specific tokens that specify the target language during generation. We use the NLLB-MD dataset (Guzmán et al., 2019; Goyal et al., 2022) in all our experiments. We use the provided split of parallel train-, dev- and test-set with respective sizes of 6000, 1310 and 1600 sentences. More on hyperparameters and implementation details are in appendix A. Our study focuses on four languages: Central Aymara, Friulian, Wolof and Russian. We use GoldFish (Chang et al., 2024) monolingual model family to score the fluency of the forward translations for every language in addition to chrF++.

First, we compare the base and trained models on these four languages using chrF++ and additionally translation fluency using Goldfish log-probabilities (Chang et al., 2024). Next, we present reward trajectories for forward and backward translation directions. Also, we present a reward function abalation study. Finally, we provide a qualitative analysis of model outputs on examples that are not seen in the training data.

5 Results

5.1 Forward Translations Evaluation

Translation Quality with chrF++ Following the original NLLB work (Goyal et al., 2022) in terms of low-resource language translation evaluation, Table 1 summarizes chrF++ scores on the forward translations before and after training. Across all languages and both model sizes, round-trip reinforcement training improves translation quality and fluency, yielding consistent absolute gains. We generally avoid model-based metrics like COMET (Rei et al., 2020) or BERTScore (Zhang et al., 2020) for forward translations due to the lack of reliable scoring models for many of the languages we evaluate. However, for Russian—where a reliable COMET model is available—we additionally report COMET scores, which change from 0.84 before training to 0.86 after training.

	600M		1.3B	
Language	Before	After	Before	After
Central Aymara	24.38	27.96 (+3.58)	26.01	28.70 (+2.69)
Friulian	45.75	47.50 (+1.75)	49.41	50.74 (+1.33)
Wolof	20.73	23.78 (+3.05)	24.73	26.81 (+2.08)
Russian	48.80	52.68 (+3.88)	50.10	54.21 (+4.11)

Table 1: Forward Translation chrF++ scores on the test set before and after training on each language for the 600M and 1.3B models using the composite reward function (Bleu + chrF++). Green values indicate absolute improvement in parentheses.

Language	Before	After	Δ
Central Aymara	-20.03	-19.87	0.16
Friulian	-19.23	-19.09	0.14
Russian	-15.64	-15.37	0.27
Wolof	-18.26	-18.07	0.19

Table 2: Goldfish model log-probability scores (natural log) before (vanilla) and after training, with improvement Δ = trained − vanilla.

The largest gain is observed on Russian (+3.88), suggesting that the intrinsic round-trip signal strongly improves adequacy and consistency even when the target language is comparatively well supported in multilingual pretraining. It's also evident that the more knowledge the model has on the language the better it is able to exploit it with the chrF++ reward signal.

For the **1.3B** model, we observe gains of **+1.33** on Friulian, **+2.08** on Wolof, **+4.11** on Russian, and **+2.69** on Central Aymara. Notably, the largest improvement again occurs on Russian (+4.11), indicating that increased model capacity does not saturate the benefits of our self-consistency objective; instead, larger models appear to better exploit the reward signal.

Translation Fluency Since we only reward the round-trip translation, it might be possible that the model generates nonsensical forward translation text, despite having an improved translation quality according to chrF++. To measure this, we check the fluency of the generated translation under the *Goldfish* language model, reporting average log-probability (natural log) before and after training (Table 2). Across all four languages, training increases the log-probability (i.e., makes it less negative), consistent with more fluent and model-preferred output distributions, confirming that our method do not damage the fluency.

The most significant improvement is again observed in Russian, mirroring the chrF++ trends.

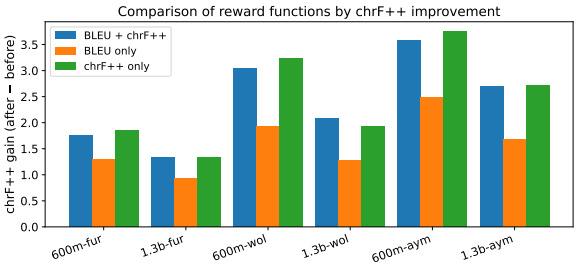

Figure 2: forward translations chrF++ gain (after – before) for different reward functions across model/language settings.

This suggests that the round-trip reinforcement objective enhances not only adequacy, as captured by chrF++, but also surface-level well-formedness and overall likelihood according to an external scorer. These results are further supported by the qualitative analysis of the translation output discussed in a later section.

5.2 Validation Curves For Forward Translations

Figure 3 shows the chrF++ scores curves on the out-of-distribution evaluation set of NLLB-MD for the four languages and both the 1.3B and 600M NLLB models. In all four languages, both models improve steadily over the course of training. The curves are smooth and approximately monotonic, without large oscillations, which indicates that the optimization procedure is stable. Across all languages and throughout training, the 1.3B model consistently outperforms the 600M model. The 1.3B curves start from higher chrF++ scores and maintain a similar margin over the 600M curves at every step. The gap remains roughly stable over time, which implies that scaling model size gives better improvements. Also, The overall shape of the curves is similar across languages, but the absolute gains differ. Central Aymara and Wolof (top-left and bottom-right) start from relatively low chrF++ scores, and both models show substantial gains over the 8K steps. This indicates that continued training is particularly helpful for low-resource languages where the baseline is under-trained. Friulian and Russian (top-right and bottom-left) have higher initial chrF++ scores and still benefit from continued training, but the gains are smaller, especially for Friulian.

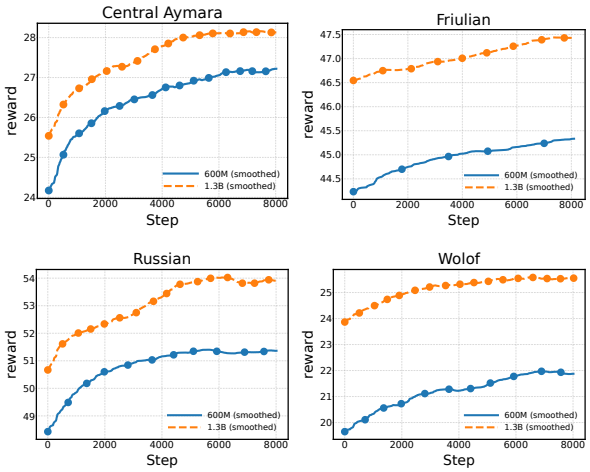

Figure 3: Validation curves showing forward translation chrF++ scores for the four languages after 8K optimization steps.

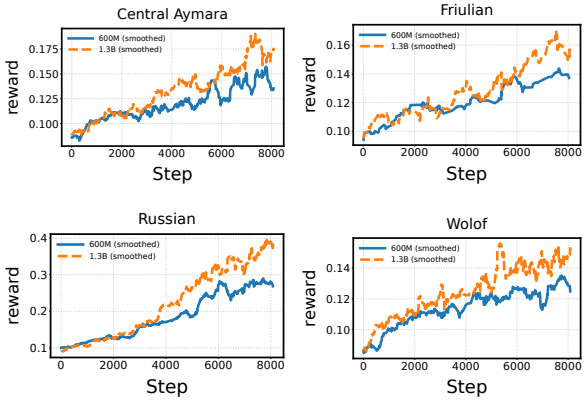

Figure 4: Training curves showing the English backward translation scores during training for the 1.3B and 600M parameter models on the 4 Languages.

Target	Before			After			ΔF1
	P	R	F1	P	R	F1	
Central Aymara	0.48	0.49	0.48	0.54	0.57	0.55	+0.07
Friulian	0.78	0.73	0.75	0.78	0.76	0.77	+0.02
Russian	0.77	0.69	0.73	0.80	0.77	0.79	+0.06
Wolof	0.53	0.48	0.51	0.54	0.55	0.55	+0.04

Table 3: Mean Precision/Recall/F1 for vanilla vs. trained models English Back-Translations.

5.3 Reward Function Ablation Study

We next study how the choice of intrinsic reward impacts training results. We compare three reward functions: (i) a uniform weighted combination of chrF++ and BLEU ("BLEU+chrF++"), (ii) BLEU only, and (iii) chrF++ only. In all cases, we keep the training procedure and hyperparameters fixed and report the *forward translation* chrF++ score before and after adaptation on each language and model size.

Figure 2 summarizes the chrF++ *gain* (after minus before). Across all six settings, BLEU-only rewards consistently improve performance but yield substantially smaller gains than rewards that include chrF++. Averaged over all conditions, BLEU-only increases chrF++ by +1.60, whereas BLEU+chrF++ yields +2.41 and chrF++-only yields +2.47.

This trend is especially pronounced for the most challenging pairs (e.g., `wol` and `aym`), where BLEU-only rewards provide a weaker learning signal under sparse n-gram matching. By contrast, chrF++ provides a smoother character-level similarity objective that appears better aligned with low-resource and morphologically rich settings, leading to the strongest and most consistent improvements (best in 4/6 conditions). Adding BLEU to chrF++ is competitive and occasionally beneficial (notably for `1.3b-wol`), but does not uniformly outperform chrF++ alone, suggesting that word-level precision signals can sometimes be redundant or slightly misaligned with the round-trip consistency objective.

5.4 Target-to-English Back-Translation Evaluation

Figure 4 shows smoothed backward-translation rewards (target → English) over training steps for the 600M and 1.3B models. As in the forward direction, both models exhibit a clear upward trend across all four languages, indicating that continued training consistently improves backward translation quality rather than leading to early overfitting. The curves are noisier than in the forward setting, which is expected given the higher variance of the backward reward signal, but the overall trajectory is monotonic: rewards increase steadily over the first several thousand steps and then continue to grow more slowly toward the end of training. Across all languages, the 1.3B model (orange) either matches or exceeds the performance of the 600M model (blue) throughout training. The larger model typically starts from slightly higher rewards and gradually widens the gap as training progresses, especially for Russian and Wolof. This pattern suggests that scaling model capacity yields consistent benefits for backward translation, and that the adaptation procedure is able to exploit the additional capacity rather than saturating at the smaller model's performance level.

Output Source	Language	chrF++	Text
Source (English)	English	N/A	That's so awesome! Which one is it?
Reference	Central Aymara	N/A	¡Ukax wali musphkayawa! ¿Kawkïrisa?
Base Model	Central Aymara	35.32	¡Wali sumäskiwa ! ¿Kawkïris uka?
Trained Model	Central Aymara	64.09	¡Ukax wali musparkañawa ! ¿Kawkïris ukaxa?
Source (English)	English	N/A	However, it was also of great importance "for a country from which we are separated by only two nations," she said.
Reference	Friulian	N/A	Dut câs, e jere ancje di grande impuartance "par une nazion de cuâl o sin separâts dome di dôs nazions," e à dite.
Base Model	Friulian	16.61	"O vin di fâ alc par fâ cressi la nestre identitât", e à dite la ministra.
Trained Model	Friulian	70.46	Dut câs, e jere ancje di grande impuartance "par un païs di dulà che o sin separâts dome di dôs nazions", e à dit jê.
Source (English)	English	N/A	Rheumatic illnesses in particular rheumatic arthritis and ankylosing spondylitis (so called Bechterew's disease)
Reference	Wolof	N/A	Jàngoroy Rheumatik rawatina arthritis rheumatik ak ankylosing spondytilis (nu duppe ko jàngoro Bechterew)
Base Model	Wolof	5.84	Jàmm yuy néew-ji-doole yuy Jàmm yuy néew-ji-doole yuy
Trained Model	Wolof	42.22	Yàmm yu réwmatik, rawatina aartritis réwmatik ak spondylitis ankylosing (yite bi ñuy wax jàngoroy Bechterew)
Source (English)	English	N/A	Sounds great to me, I'm so excited! What is the horse's name?
Reference	Russian	N/A	Zvuchit zdorovo, ya tak rad! Kak zovut konya?
Base Model	Russian	21.32	- Kak zovut loshadey?
Trained Model	Russian	74.52	Zvuchit zdorovo, ya tak rada! Kak zovut loshadey?

Table 4: Example outputs for Source (English), Reference translations, Base Model, and Trained model across four languages. Tokens in green match the reference, tokens in red differ from the reference, and tokens in blue mark segments where the trained model is closer to the reference than the base model. chrF++ scores are shown for model outputs with respect to the Reference sentence.

BERTScore To validate our findings beyond chrF++ and bleu score in which are used as a reward function, we also show the improvement with BERTScore (Zhang et al., 2020). Table 3 shows the BERTScores on the back-translated English sentences before and after training. Across all four target languages, the trained model improves BERTScore F1 relative to the vanilla baseline, with gains ranging from +0.02 (Friulian) to +0.07 (Central Aymara). The improvements are generally driven by higher recall after training (notably Central Aymara and Russian), while precision also increases or remains stable, indicating the trained back-translations are both more complete and at least as faithful under this semantic metric. Our findings show consistent improvement upon different metrics.

5.5 Qualitative Analysis On Translation Outputs

Finally, in Table 4 we provide a qualitative comparison of translations produced by the 1.3B base model and the trained model across four evaluation languages (one example each; additional samples are reported in Appendix B). Overall, the trained model produces outputs that are close to the human reference.

For **Central Aymara**, the base model captures the positive tone but diverges in lexical choice and phrasing. The trained model matches the reference opener ("*Ukax wali . . .*") and better recovers the intended predicate meaning, yielding a question form that is closer to the reference. This suggests that training primarily improves lexical selection and common conversational templates in this low-resource setting.

For **Friulian**, the base model fails to translate the source sentence and instead generates unrelated quoted text (English: "We have to do something to grow our identity," the minister said.). In contrast, the trained model restores the intended discourse structure and recovers most of the reference content ("*Dut câs, e jere ancje di grande impuartance . . .*"), with only a minor mismatch in the final pronoun choice and the final translation to English is faithful to the reference. The chrF++ jump (16.61 → 70.46)

indicates substantially improved faithfulness and reduced semantic drift.

For **Wolof**, the baseline output degenerates into repetition, a common failure mode for low-resource and domain-specific inputs. The trained model largely recovers the technical terminology and coordination structure (e.g., *aartritis réwmatik*, *spondylitis ankylosing*, and the parenthetical *Bechterew*), though some surrounding function words remain imperfect. The corresponding chrF++ gain (5.84 $\rightarrow$ 42.22) suggests improved robustness (avoiding repetition) and better retention of specialized terminology.

Finally, for **Russian**, the base model produces a partially correct question but uses an incorrect lexical/inflectional form for "horse." The trained model matches the reference more closely in the preceding clause and improves morphosyntactic realization (e.g., correct gender marking in "*rada*"), while the noun error persists. The chrF++ increase (21.32 $\rightarrow$ 74.52) highlights improved fluency and morphology, with remaining challenges in precise lexical choice.

Taken together, these examples suggest that self-supervised training improves both (i) content coverage and faithfulness (especially evident in Friulian and Wolof) and (ii) surface-form realization, including idiomatic openers and inflectional morphology (Central Aymara and Russian). Residual mismatches point to remaining difficulties in fine-grained lexical selection and domain-specific function-word control. Nevertheless, despite relying only on monolingual reward signals, the model is able to self-improve on the target low-resource languages without any direct supervision.

6 Conclusion

We presented a self-supervised RL fine-tuning approach for low-resource machine translation based on round-trip bootstrapping with NLLB. By translating English into a target language and rewarding the quality of the reconstructed English using intrinsic signals (chrF++ and BLEU), GRPO training yields consistent chrF++ improvements on an out-of-distribution test set for Central Aymara, Friulian, Wolof, and Russian across both 600M and 1.3B models. Qualitatively, the trained models reduce common low-resource failure modes such as repetition and semantic drift, and produce outputs that are more fluent and faithful to the reference.

Our ablations show that chrF++ provides the most reliable learning signal in this setting, while adding BLEU is competitive but not uniformly better. Overall, these results indicate that, even with only monolingual reward signals, pretrained multilingual models can self-improve translation quality for low-resource languages without direct parallel supervision.

7 Limitations

Our study has several limitations that point to clear directions for future work.

Scaling to larger NLLB checkpoints. Due to resource constraints, we only evaluate the distilled 600M and 1.3B NLLB variants, and do not test whether the same round-trip GRPO training remains stable and beneficial for substantially larger models (e.g., 3.3B NLLB). This is because empirical scaling laws suggest that multilingual MT quality often improves predictably with model scale, and that scaling can change the effective capacity allocated to individual language pairs, including out-of-distribution behavior (Fernandes et al., 2023; Kaplan et al., 2020).

Broader low-resource language coverage within NLLB-MD. We focus on four languages (Central Aymara, Friulian, Wolof, and Russian) and do not evaluate on the remaining low-resource languages included in NLLB-MD (notably Bhojpuri and Dyula) (Goyal et al., 2022). Extending experiments to these additional languages would strengthen conclusions about robustness across scripts, typology, and domain mismatch, and would help identify when round-trip rewards are most effective versus when they induce overly paraphrastic solutions.

Richer reward functions for low-resource MT. Our intrinsic reward is a simple linear combination of chrF++ and BLEU computed on the reconstructed English, which is effective but limited. Future work should explore richer and more robust reward design, for example: (i) adding learned semantic metrics (e.g., COMET/BLEURT-style signals) to reduce reliance on lexical overlap, (ii) using target-side fluency rewards (monolingual LM scoring) to better control grammaticality.

8 Ethical Considerations

Our approach fine-tunes multilingual MT models using self-supervised, round-trip reinforcement learning (English$\rightarrow$target$\rightarrow$English) with au-

tomatic rewards (chrF++/BLEU). While this avoids collecting new human-labeled parallel data, it introduces several potential risks:

Amplification of model errors and biases. Because training relies on the model's own generations, systematic errors (e.g., gender/role stereotypes, offensive terms, dialectal mismatches) can be reinforced. Biases present in pretraining or in monolingual English data may propagate into low-resource outputs.

Hallucinations and harmful content translation. RL may increase fluency while still hallucinating facts, altering named entities, or mistranslating safety-critical content (medical/legal). Additionally, improved MT quality could facilitate translation of toxic or hateful content into additional languages.

Mitigations. We mitigate these risks by (i) monitoring for degeneracy (repetition/copying) and reporting qualitative examples, (ii) evaluating with multiple metrics (chrF++, BLEU, BERTScore on back-translations, and target-side fluency), (iii) limiting training data to vetted corpora and documenting compute settings for transparency.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Preprint*, arXiv:1506.03099.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *Preprint*.

Alexis Conneau and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. *ACL*.

Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China. Association for Computational Linguistics.

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. *Preprint*, arXiv:2504.10160.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10053–10071. PMLR.

R. A. Fisher. 1992. *Statistical Methods for Research Workers*, pages 66–70. Springer New York, New York, NY.

Naman Goyal and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.

Barry Haddow and 1 others. 2022. Survey of machine translation for low-resource languages. *Computing Research Repository*.

Jianfei He, Shichao Sun, Xiaohua Jia, and Wenjie Li. 2024. Recovery should never deviate from ground truth: Mitigating exposure bias in neural machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 68–79, Sheffield, UK. European Association for Machine Translation (EAMT).

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Xiuming Huang. 1990. A machine translation system for the target language inexpert. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. *Preprint*, arXiv:1711.00043.

Andrii Maksai and Pascal Fua. 2018. Eliminating exposure bias and loss-evaluation mismatch in multiple object tracking. *Preprint*, arXiv:1811.10984.

Paul McNamee and Kevin Duh. 2023. An extensive exploration of back-translation in 60 languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8166–8183.

Sonja Nießen, Franz Josef Och, Gregor Leusch, Hermann Ney, and 1 others. 2000. An evaluation tool for machine translation: Fast evaluation for mt research. In *LREC*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

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.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. *Preprint*, arXiv:1511.06732.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. *Preprint*, arXiv:1511.06709.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*, second edition. The MIT Press.

Ashish Vaswani and 1 others. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Weiyue Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552.

Hyperparameter	**Value**
Learning rate	2×10^{-6}
Batch size	2
Group size K	4
Generation temperature	1.8
Top-k sampling	100
Top-p sampling	0.95
KL coefficient β	0.04
Clip parameter ϵ_c	0.2
Reference update frequency	16 steps
Maximum sequence length	64 tokens
Training epochs	2

Table 5: Hyperparameters used for training.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. *Preprint*, arXiv:1808.08866.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Implementation Details

We implement our method using PyTorch (Paszke et al., 2019) and the Hugging Face Transformers (Wolf et al., 2020) library. Table 5 summarizes the key hyperparameters used in our experiments.

We use AdamW (Kingma and Ba, 2017) for optimization. During generation, we employ nucleus sampling with temperature scaling to encourage diverse candidate generation.

B Qualitative Translation Examples

B.1 Central Aymara

Reference: Ukax jichhax amuyt'añjamaxïwa, kipka irnaqawinakjamakiwa. Jaqinakax kunti munapki ch'amañchañawa. ¡kayak anatañ lurañ yant'äsma! Wali jasakiw apnaqañaxa, qutan samart'añjamaspawa jan ukasti qhuru umanakan

chuyma jalsuyasirjamarakispawa.
Source (EN): Ah that makes a ton of sense, they seem like similar jobs at least in some respects. Encouraging people and helping them meet their goals. You should totally try kayaking! It's super versatile, it can be relaxing on a lake or thrilling doing whitewater.
Vanilla (system output): Ah, al à sens, a somein lavoris similârs almancul in cualchi sens. incoragjâ lis personis e judâlis a rivâ ai lôr obietîfs. Tu varessis di provâ il kayak! al è une vore versatile, al pues jessi un lavôr di
Vanilla (EN gloss): Ah, al à sens, a somin lavoris similarârs at least in some sens. incoragjâ lis personis e judâlis a rivâ ai lôr obietîfs. Tu varessis di provâ il kayak! al è une vore versatile, al pues jessi un lavôr di
Trained (system output): Ah, al à sens, a somein lavoris similis almancul in cualchi sens. Incoragjâ lis personis e judâlis a rivâ ai lôr obietîfs. Tu varessis di provâ a fâ il kayak! Al è une vore versatile, al pues jessi ri
Trained (EN gloss): Ah, in a sense, some people work similarly at least in some senses. Incoragjâ the people and Jews to achieve their objectives. You have a variety of ways to try to do the kayaking! Al è une vore versatile, you can be very versatile, and you can be the best in the world.

Reference: O ai apene sfornât i biscos PB cu la forme dai mei cjans!
Source (EN): I baked PB cookies in the shape of my dogs!
Vanilla (system output): O ai fat biscots di PB a forme dai miei cans!
Vanilla (EN gloss): Or PB biscuits shaped like my dogs!
Trained (system output): O ài fat biscuts PB a forme dai miei cans!
Trained (EN gloss): I made PB cookies in the shape of my dogs!

Reference: O soi sigur che al divente mat, ma tu asreson, il mangja co cusini no l'è just par lôr. Perciò o ten simpri sot man al di bon par lôr quant co cusini.
Source (EN): I am sure the smells drive him crazy, but you are right. The food I bake is not good for them. I have dog treats on hand anytime I bake for this reason.
Vanilla (system output): O soi sigure che i odôrs lu fasin deventâ mat, ma tu âs reson. Il mangjâ che o fâs no je buine par lôr. Par chest motîf o ai simpri lis cjasis a man.
Vanilla (EN gloss): I'm sure the smells drive them crazy, but you're right. The food I make isn't good for them. That's why I always keep the houses handy.
Trained (system output): O soi sigure che i odôrs lu fasin deventâ mat, ma tu âs reson. Il mangjâ che o fâs jo nol è bon par lôr. Par cheste reson o ai simpri lis cjampanis pai cans in man.
Trained (EN gloss): I am sure the smells make them go crazy, but you are right. The food I make is not good for them. That is why I always have the dogs' bells in my hand.

B.2 Friulian

Reference: O ai apene sfornât i biscos PB cu la forme dai mei cjans!
Source (EN): I baked PB cookies in the shape of my dogs!
Vanilla (system output): O ai fat biscots di PB a forme dai miei cans!
Vanilla (EN gloss): Or PB biscuits shaped like my dogs!
Trained (system output): O ài fat biscuts PB a forme dai miei cans!
Trained (EN gloss): I made PB cookies in the shape of my dogs!

Reference: O soi sigur che al divente mat, ma tu asreson, il mangja co cusini no l'è just par lôr. Perciò o ten simpri sot man al di bon par lôr quant co cusini.
Source (EN): I am sure the smells drive him crazy, but you are right. The food I bake is not good for them. I have dog treats on hand anytime I bake for this reason.
Vanilla (system output): O soi sigure che i odôrs lu fasin deventâ mat, ma tu âs reson. Il mangjâ che o fâs no je buine par lôr. Par chest motîf o ai simpri lis cjasis a man.
Vanilla (EN gloss): I'm sure the smells drive them crazy, but you're right. The food I make isn't good for them. That's why I always keep the houses handy.
Trained (system output): O soi sigure che i odôrs lu fasin deventâ mat, ma tu âs reson. Il mangjâ che o fâs jo nol è bon par lôr. Par cheste reson o ai simpri lis cjampanis pai cans in man.
Trained (EN gloss): I am sure the smells make them go crazy, but you are right. The food I make is not good for them. That is why I always have the

dogs' bells in my hand.

Reference: No tant. Stât impegnât a ciri plas gnûs. Soi sigû co tacarai a vore a mae quat co sarai stac di cusinà.
Source (EN): Not recently. I have been consumed by trying to find recipes. I tend to hobby jump quite a bit. I am sure I will start knitting like crazy once I get tired of baking.
Vanilla (system output): No dome di pôc, mi soi consumade a cirî ricetis, o soi propit un pôc a saltâ fûr di hobby, o soi sigure che o fasarai il tricot come une matete cuant che mi stufarà di fâ il pan.
Vanilla (EN gloss): Not only that, I've been consumed with finding recipes, or I've been kind of jumping out of my hobby, or I'm sure I'll be knitting like a mat when I get tired of baking.
Trained (system output): No di pôc timp. O soi stât consumât a cirî ricetis. O ten a saltâ un grum di hobbies. O soi sigûr che o tacarà a fâ il tricot come un mat cuant che o varès stufât di fâ il pan.
Trained (EN gloss): Not recently. I have been consumed looking for recipes. I want to skip a lot of hobbies. I am sure that I will start knitting like crazy when I get tired of baking bread.

B.3 Russian

Reference: Da, no ya podumyvayu provesti vikend v etom kan'yone ryadom so mnoy. Eto ne plyazh, no i ne les, dolzhno byt' zdorovo.
Source (EN): I do, but I am thinking of doing a weekend in this canyon near me. It's not the beach, but it's nto a forest either, should be cool.
Vanilla (system output): Ya dumayu provesti vykhodnye v kan'one nedaleko ot menya.
Vanilla (EN gloss): I'm thinking of spending the weekend in a canyon near me.
Trained (system output): Da, no ya dumayu provesti vykhodnye v etom kan'one ryadom so mnoy. Eto ne plyazh, no eto tozhe les, dolzhno byt' kruto.
Trained (EN gloss): Yes, but I'm thinking of spending the weekend in that canyon next to me. It's not the beach, but it's also the forest, so that should be cool.

Reference: Tochno. Kak tvoya sem'ya? Skol'ko vas?
Source (EN): That's right. How is your family? how many of you are there?
Vanilla (system output): Kak vasha sem'ya?
Vanilla (EN gloss): How's your family?
Trained (system output): Tochno. Kak vasha sem'ya? Skol'ko vas tam?
Trained (EN gloss): Exactly. How is your family? How many are you there?

Reference: Nikakikh problem! Skol'ko i kakikh?
Source (EN): No problem at all! How many and which ones?
Vanilla (system output): – Skol'ko i kakie?
Vanilla (EN gloss): - How many and what?
Trained (system output): Nikakikh problem! Skol'ko i kakie?
Trained (EN gloss): No problem! How many and what?

B.4 Wolof

Reference: Kuréelu réewum Suède bu wérguyaram (Rättsmedicinalverket) tàmbali woon na di def tests yi ci lu teel ci at mi.
Source (EN): Sweden's national Forensic Medicine Agency (Rättsmedicinalverket) started carrying out the tests earlier year.
Vanilla (system output): Bësum-bésum-bésum-bésum-bésum-bésum-bésum-bésum-bésum-bésum-bésum-bésum-bésum-bés
Vanilla (EN gloss): Day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day-day
Trained (system output): Bànqaasu Farãs gu réew mi ci Suweed (Rättsmedicinalverket) tàmbalee jëfandikoo ay seetlu ci njëlbéenug at mi.
Trained (EN gloss): The French national bank of Sweden (Rättsmedicinalverket) began using the studies at the beginning of the year.

Reference: Moom ak mbiipkat bu dëkk Bulgari Miroslav Petkov, dafay def ab konseer yu am bàkk biiG̈éej gu ñuul gi tasena ak Àll bu ñuul.;, wala sax, B̈alkan folkloreB̈aje naak ab njàng mu Allëmaañ bu kalasik ak taxawaay bu dëggër ci misiku brass.
Source (EN): Together with the Bulgarian trumpeter Miroslav Petkov, he is giving a concert that carries the slogan "The Black Sea meets the Black Forest", or even, "Balkan folklore" meets a classic German education with roots in brass music.
Vanilla (system output): Ci biir li muyànd ak mbëjug buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum buum
Vanilla (EN gloss): In addition to its use as an anti-inflammatory, it is also used as an anti-

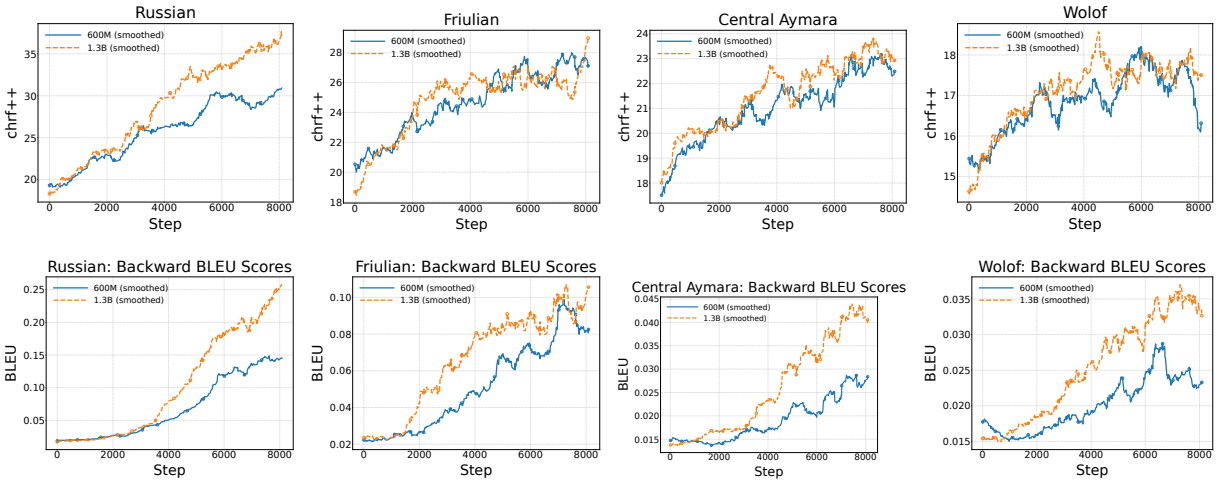

Figure 5: Validation curves after 8K optimization steps. Top row: forward translation chrF++ scores. Bottom row: backward translation BLEU scores.

inflammatory and anti-inflammatory agent.
Trained (system output): Ci biir àndandoo ak mbëjug buulgar bi di Miroslav Petkov, mu ngi doon jox ab kaaraange bu am kàdduy waxtaan bi di "Géej gu Màgg gi daje ak Géej gu Màgg gi", walla sax, "Balkan folkl
Trained (EN gloss): In collaboration with the Bulgarian writer Miroslav Petkov, she was developing a cover with the words of the speech "The Old Sea meets the Great Sea", or even, "Balkan folklore"

Reference: Ñeen lañu! Amna sama benn makk bu goor ak samay waajur. Ñun goor yi deñoo bëgg bannexuloo ci benn sigaar ci ngoon gi.
Source (EN): There are four of us! I have an older brother and my two parents. Us men like to enjoy a cigar together in the evenings.
Vanilla (system output): Am nanu ñeenti nit, am naa magam ak samay maam ak sama baay.
Vanilla (EN gloss): We had four children, one of whom died in infancy.
Trained (system output): Am nanu ñeenti nit! Am naa mag bu góor ak samay ñaari waajur. Nun góor yi, bëgg nanu a lekk cigare ci ngoon ci ngoon.
Trained (EN gloss): We have four people! I have a older brother and my two parents. We men like to smoke in the afternoon in the afternoon.

C Additional Training Plots

Figure 5 reports validation curves for the forward and backward translation models over 8K optimization steps. The top row shows chrF++ for forward translation, and the bottom row shows BLEU for backward translation, for Russian (Cyrl), Friulian (Latn), Aymara (Latn), and Wolof (Latn). These curves are included for completeness and to illustrate the round-trip reinforcement learning training dynamics across languages and directions. For both the forward and backward translations during training, it's evident that there is no plateau and the models can further benefit from continued RL training.