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 approaches for improving low-resource MT remain underexplored. We investigate a self-supervised reinforcement learning 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, Dyula, Bhojpuri 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. Code available at: https://anonymous.4open.science/r/MT-via-Round-Trip-RL-D413/README.md

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 many languages continue to face limited access to digital information and reduced participation in multilingual technologies.

Recent advances in massively multilingual machine translation have produced models that cover hundreds of languages within a single system. In particular, the No Language Left Behind (NLLB) project (Team et al., 2022) introduced a large-scale multilingual translation model trained on data covering more than 200 languages, including many low-resource ones. By learning shared multilingual representations, such models enable cross-lingual transfer that improves translation quality even when direct parallel supervision is limited. Nevertheless, adapting these models to further improve low-resource directions remains challenging when parallel data is unavailable.

In this work, we explore whether pretrained multilingual translation models can improve translation quality using only monolingual signals. We introduce a self-supervised framework that combines round-trip translation with reinforcement learning (RL). Given an English sentence, the model first generates a translation into a target low-resource language and then translates the result back into English. If the translation preserves the original meaning, the reconstructed sentence should closely match the source. We therefore use similarity metrics such as BLEU and chrF++ to provide rewards that guide learning.

Standard machine translation training relies on maximum likelihood estimation (MLE) (Fisher, 1992), which predicts tokens conditioned on ground-truth histories. While effective, MLE suffers from exposure bias (Bengio et al., 2015; Wang and Sennrich, 2020; He et al., 2024) and an objective mismatch (Maksai and Fua, 2018), since training does not directly optimize evaluation metrics such as chrF++ (Popović, 2015). These issues are particularly pronounced in low-resource settings, where limited supervision makes models more susceptible to compounding errors during inference.

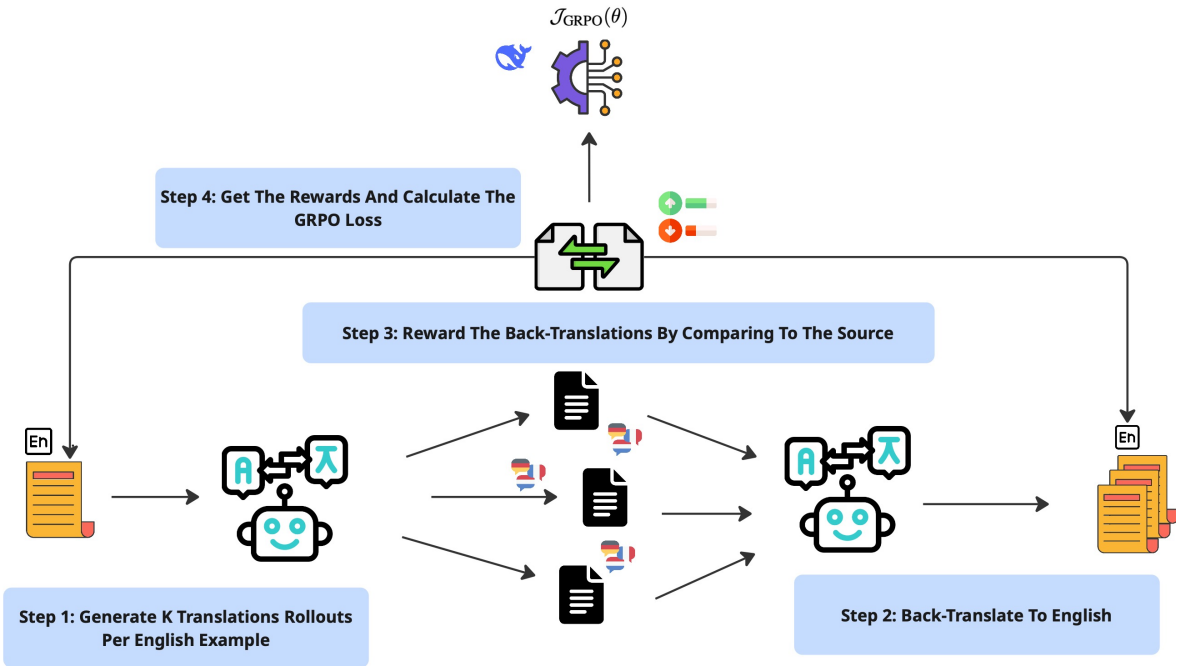

Figure 1: Overview Of Our Low-Resource Machine Translation With Round-Trip Reinforcement Learning Method.

Back-translation (Sennrich et al., 2016) is the dominant approach for leveraging monolingual data in low-resource MT. However, its effectiveness depends heavily on the quality of the reverse translation model, and errors in early translations can propagate through synthetic training data (McNamee and Duh, 2023). Moreover, standard self-supervised approaches typically rely on maximum likelihood training rather than directly optimizing translation metrics.

Reinforcement learning provides a natural framework for addressing these limitations. By treating translation as a sequential decision-making process, RL allows models to optimize non-differentiable evaluation metrics and learn from their own predictions during training (Sutton and Barto, 2018). Prior work has explored RL for neural machine translation (Ranzato et al., 2016; Wu et al., 2018), and recent studies have applied RL techniques to large language models for translation (Feng et al., 2025). However, the use of RL to improve round-trip translation consistency for low-resource languages remains relatively underexplored.

We propose a reinforcement learning approach that directly optimizes round-trip reconstruction quality using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) applied to pretrained NLLB models (Team et al., 2022). This approach allows the model to improve translation quality without requiring additional parallel data.

Our contributions are as follows:

- We introduce a reinforcement learning framework for improving low-resource translation using round-trip consistency without parallel training data.
- We demonstrate that optimizing reconstruction rewards improves translation fluency and adequacy in low-resource settings.
- We provide empirical analysis across several languages and model sizes.

2 Related Work

2.1 Improving NMT with Monolingual Data and Round-Trip Translations

Back-translation is a widely used technique for leveraging monolingual data in neural machine translation. Sennrich et al. (2016) showed that generating synthetic source sentences from target-language monolingual data can significantly improve translation quality when combined with genuine parallel corpora.

Building on this idea, unsupervised machine translation methods have demonstrated that translation systems can be trained without parallel data by combining denoising autoencoding, back-translation, and cycle-consistency objectives (Lample et al., 2018). These approaches enforce consistency between forward and backward translation

models, encouraging translations that preserve the original meaning after a round-trip process.

Round-trip translation has also been explored in other contexts. Early work by Huang (1990) used round-trip translation as a quality-checking tool for users unfamiliar with the target language. Federmann et al. (2019) later demonstrated that translating through multiple pivot languages can generate diverse paraphrases, a process known as "multilingual whispers."

Other self-supervised approaches attempt to mine parallel sentences from comparable corpora. For example, Ruiter et al. (2019) propose jointly mining sentence pairs and training translation models using cross-lingual similarity signals derived from internal model representations. Their results show that high-quality translation models can be trained from comparable data without direct parallel supervision.

2.2 Reinforcement Learning for Neural Machine Translation

Reinforcement learning has been explored as a way to optimize sequence-level objectives in machine translation. Ranzato et al. (2016) and Wu et al. (2018) showed that policy gradient methods can reduce exposure bias and directly optimize evaluation metrics. These approaches typically combine RL objectives with maximum likelihood training to stabilize optimization.

More recently, RL techniques have been applied to large language models for translation. Feng et al. (2025) propose MT-R1-Zero, which applies RL without supervised warm-starts and uses rule-based reward functions to improve translation quality. Their method employs Group Relative Policy Optimization (GRPO) (Shao et al., 2024), an algorithm derived from Proximal Policy Optimization (PPO) (Schulman et al., 2017) that removes the value critic and normalizes rewards across sampled outputs.

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) (Shao et al., 2024) with back-translation-based rewards, enabling effective adaptation without parallel data augmentation.

3.1 Preliminary

For each input q, GRPO (Shao et al., 2024) samples a group of G candidate outputs $\{o_1, \ldots, o_G\} \sim \pi_{\theta_{\text{old}}}(y \mid x)$ from the current policy. The update maximizes

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \Bigg[\\ \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\,\hat{A}_{i,t}, \\ \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon_c, 1+\varepsilon_c\right)\hat{A}_{i,t}\Big) \\ -\beta\, D_{\text{KL}}\big[\pi_\theta \| \pi_{\text{ref}}\big]\Bigg], \\ \text{where } r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}.\end{aligned} \tag{1}$$

GRPO differs from PPO in how it defines the advantages $\hat{A}_{i,t}$. Instead of subtracting a critic prediction, GRPO normalizes rewards within the sampled group. Let $R_i = R(x, y_i)$; then

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\})}. \tag{2}$$

When all or none of the candidates solve the problem, every $\hat{A}_{i,t}$ is zero and the policy gradient vanishes except for the KL term.

3.2 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.

3.3 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.

3.4 Reward Design

We employ a character-level reward function using chrF++.

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:

$$\text{chrF++} = (1 + \beta^2) \cdot \frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}} \tag{5}$$

BLEU. (Papineni et al., 2002) provides complementary word-level precision signals. We mainly use BLEU in our ablation studies.

4 Experiments

4.1 Experimental Setup

We base our experiments on NLLB-200 (No Language Left Behind) (Team 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 (Guzman et al., 2019; **?**) in all our experiments, and follow the provided split of parallel train, dev, and test sets with respective sizes of 6000, 1310, and 1600 sentences. We discard the target side during training., hence the self-supervised approach. We implement our method in PyTorch (Paszke et al., 2019) using the Hugging Face Transformers library (Wolf et al., 2020). We optimize with AdamW (Kingma and Ba, 2017) using a learning rate of 2×10^{-6} for 2 epochs with batch size 2, group size K=4, KL coefficient β=0.04, clip parameter ϵ_c=0.2, and a reference update frequency of 16 steps. During generation, we use nucleus sampling with temperature scaling (temperature 1.8, top-k 100, top-p 0.95) to encourage diverse candidate generation. Our study focuses on the following languages: Central Aymara, Friulian, Wolof, Bhojpuri, Dyula and Russian. We use the GoldFish (Chang et al., 2024) monolingual model family to evaluate the fluency of our system's forward translations for each language.

4.2 Baselines

Unsupervised Machine Translation Using Monolingual Corpora Only (round-trip / cycle training). We include Lample et al. (2018) method with denoising and cycle-consistent reconstruction objectives, augmented with an adversarial discriminator. Models are trained for 2 epochs with batch size 2. We use separate learning rates for the encoder/decoder and discriminator: $\text{LR}_{\text{enc/dec}} = 3 \cdot 10^{-5}$ and $\text{LR}_{\text{disc}} = 5 \cdot 10^{-5}$. Objective weights are set to $\lambda_{\text{auto}} = 1.0$ (denoising auto-encoding), $\lambda_{\text{cd}} = 1.0$ (cycle/round-trip reconstruction), and $\lambda_{\text{adv}} = 1.0$ (adversarial alignment). Input corruption uses word dropout 0.1 and local shuffle with window size $k = 3$.

Back-Translation We implement the standard back-translation method of Sennrich et al. (2016). Given monolingual target-language sentences y, we first generate synthetic English source sentences $\hat{x}$ using the pretrained NLLB model in the target→English direction. The resulting synthetic parallel pairs $(\hat{x}, y)$ are then used to train the English→target translation model using the standard cross-entropy objective. Synthetic translations are generated using the same pretrained NLLB model prior to reinforcement learning fine-tuning.

Table 1: Main results (higher is better). For each language and model size, we **bold** the best score and underline the second-best across {Base, BT, UMT, RT, Ours}. BT denotes target-to-source back-translation (tgt→src). UMT (Unsupervised Machine Translation with Monolingual Corpora Only), RT (round-trip), and Ours are trained via round-trip translation (src→tgt→src). We set $\lambda_{\text{chrF}} = 1$ and $\lambda_{\text{BLEU}} = 0$.

	NLLB-600M					NLLB-1.3B				
Language	Base	BT	UMT	RT	Ours	Base	BT	UMT	RT	Ours
Central Aymara (ayr_Latn)	24.38	27.34	24.85	26.68	**28.13**	26.01	28.61	28.21	27.59	**28.73**
Wolof (wol_Latn)	20.73	23.57	19.89	22.84	**23.96**	24.73	26.35	**27.72**	25.21	26.66
Russian (rus_Cyrl)	48.80	**53.31**	51.70	51.63	52.79	50.10	**55.46**	55.14	54.07	55.16
Friulian (fur_Latn)	45.75	47.38	**48.96**	46.25	47.61	49.41	**51.72**	46.43	49.68	50.76
Dyula (dyu_Latn)	18.03	20.13	20.82	21.89	**22.50**	18.24	20.89	22.93	21.96	**22.97**
Bhojpuri (bho_Deva)	43.07	**45.36**	39.79	41.62	43.97	44.31	**47.23**	44.43	44.52	44.77
Avg.	34.91	36.18	34.34	35.15	**36.49**	37.56	**38.38**	37.48	37.17	38.18

Round-Trip Translation with MLE Finally, a round-trip translation baseline (van Zaanen and Zwarts, 2006) where sentences are translated from English to the target language and then translated back to the English using the same model parameters. The model is trained to reconstruct the original sentence after the two-step translation process using standard cross entropy objective.

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
Bhojpuri	-14.93	-14.97	-0.04
Dyula	-17.53	-16.43	1.09

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

5 Results

5.1 Forward Translations Evaluation (English->Target)

Translation Quality with chrF++ Following the original NLLB work (Team 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.

Our method is particularly effective for very low-resource languages. With the 600M model, it achieves the best scores on Central Aymara (28.13), Wolof (23.96), and Dyula (22.50), outperforming BT by 0.79, 0.39, and 2.37 points respectively. It also consistently exceeds the round-trip (RT) baseline on these languages. With the 1.3B model, our method remains best on Central Aymara (28.73) and Dyula (22.97), while remaining competitive on Wolof (26.66), where UMT attains the top score (27.72). For higher-resource languages (Russian, Friulian, Bhojpuri), BT is typically the strongest baseline, but the gap to our method remains small. Importantly, BT relies on target-to-English translation during training (tgt→English), effectively assuming access to a reverse-direction model or parallel signal. In contrast, our method only assumes the availability of an English monolingual corpus and does not require target→English supervision. Despite this weaker assumption, our approach remains competitive with BT for very low-resource languages like Dyula, Central Aymara and Wolof. Notably, our method consistently improves over the vanilla round-trip (RT) baseline across all languages and model sizes. On average, our method achieves the best performance with the 600M model (36.49), outperforming BT by 0.31 points. With the 1.3B model, it remains competitive (38.18), within 0.20 of the strongest method (BT at 38.38). Finally, UMT exhibits less consistent behavior across languages and model sizes, which is plausibly explained by its more complex optimization objective that combines auto-encoding, cross-domain, and adversarial losses.

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 measure target-side fluency using a monolingual LM under the *Goldfish* Chang et al. (2024) language model, reporting average log-probability (natural log) before and after training (Table 2). Across five of the six languages, training improves LM likelihood (i.e., makes it less negative), consistent with more fluent and model-preferred output distributions, confirming that our method does not degrade fluency. For Bhojpuri, we think that due to the higher morphological variability in this language, the system does not improve fluency.

These scores suggest 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.

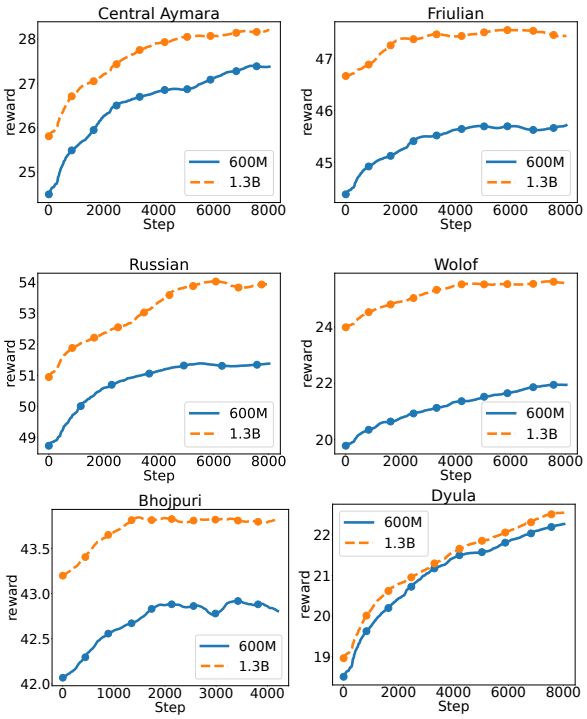

Figure 2: Validation curves showing forward translation chrF++ scores (reward) for the six languages after 8K optimization steps.

5.2 Validation Curves For Forward Translations (English->Target)

Figure 2 shows the chrF++ scores curves on the out-of-distribution evaluation set of NLLB-MD for the six languages and both the 1.3B and 600M NLLB models. In all six 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.

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 3 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., `Wolof` and `Central Aymara`), 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. Finally, as shown in figure 4, we experimented with BLEURT (Sellam et al., 2020) as the reward in the English recon-

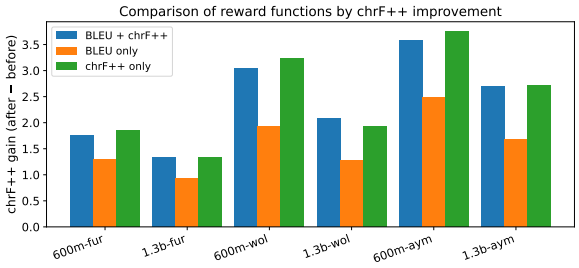

Figure 3: forward translations chrF++ gain (after – before) for different reward functions across model/language settings. (fur=Friulian, wol=Wolof, aym=Central Aymara and rus=Russian)

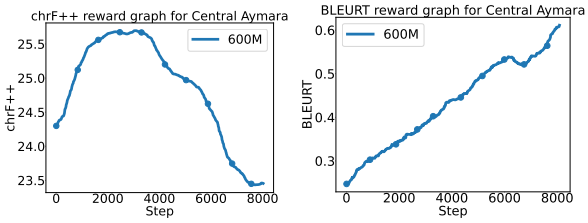

Figure 4: Results after training the 600M model with BLEURT as the main back-translation reward on Central Aymara. Left: chrF++ reward graph on English->Target direction on the validation set. Right: BLEURT Scores during training between the source and generated English back-translation.

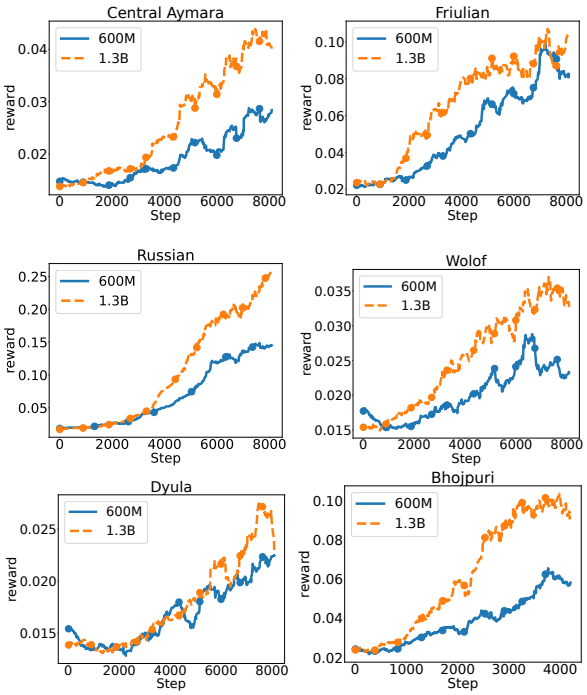

Figure 5: Training curves showing the tgt->English backward translation BLEU scores (reward) during training for the 1.3B and 600M parameter models for the 6 Languages across 8K optimization steps.

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
Dyula	0.47	0.35	0.40	0.56	0.56	0.56	+0.16
Bhojpuri	0.80	0.77	0.79	0.77	0.77	0.77	-0.02

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

struction step, comparing reconstructed sentences to the original English source. However, training became unstable: the model rapidly exploited the learned metric, producing outputs that achieved high BLEURT scores while true translation quality deteriorated, leading to a collapse in forward translations chrF++ scores on the validation set. This behavior is consistent with reward over-optimization (Paulus et al., 2017; Lazaridou et al., 2020) when using learned metrics as training objectives.

5.4 Back-Translation Evaluation (Target->English)

Figure 5 shows backward-translation rewards (target → English) over training steps for the 600M and 1.3B models. Both models exhibit a clear upward trend across all six 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. It is clear 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. Finally, this aligns with the forward translation reward graphs and our hypothesis is sound: RL optimization for the roundtrip back-translation chrF++ scores improves translation in both directions.

BERTScore To validate our findings beyond chrF++ and BLEU, which are used as reward signals during training, we also report results using BERTScore (Zhang et al., 2020). Table 3 shows the BERTScores computed on the back-translated English sentences before and after training. The trained model improves BERTScore F1 in five of the six cases, with gains ranging from +0.02 (Friulian) to +0.16 (Dyula). The largest improvements occur for Dyula (+0.16). The decrease in Bhojpuri again (-0.02) can be due to weaker bilingual

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 paîs 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.

alignment and a higher morphological variability. However, the decrease is very small. The substantial gain for Dyula reflects improvements in both precision and recall, suggesting that training yields back-translations that are both more semantically complete and better aligned with the reference under this embedding-based metric.

5.5 Qualitative Analysis On Translation Outputs

Table 4 presents a qualitative comparison between the 1.3B base model and the trained model in four evaluation languages (one example each; additional samples appear in Appendix A). Overall, the trained model produces outputs that are substantially closer to the human reference. The examples indicate that self-supervised training improves both (i) content coverage and faithfulness (notably in Friulian and Wolof) and (ii) surface realization, including idiomatic openers and morphology (Central Aymara and Russian). The remaining errors primarily involve fine-grained lexical choice and function-word control. Despite relying only on monolingual reward signals, the model improves translation quality in low-resource languages without direct supervision.

6 Conclusion

We propose a self-supervised RL fine-tuning approach for low-resource machine translation using round-trip bootstrapping with NLLB. The model translates English into a target language and is rewarded based on the quality of the reconstructed English using intrinsic metrics (chrF++ and BLEU). GRPO training yields consistent chrF++ improvements on an out-of-distribution NLLB-MD test set for both 600M and 1.3B models. Qualitatively, the trained models reduce common low-resource errors such as repetition and semantic drift, producing outputs that are more fluent and faithful. Ablations show chrF++ provides the most reliable learning signal, while combining BLEU is competitive but not consistently better. These results suggest pretrained multilingual models can self-improve low-resource translation using only monolingual reward signals without 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). 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 six languages and thus extending experiments to 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.

8 Ethical Considerations

Our approach fine-tunes multilingual MT models using self-supervised, round-trip reinforcement learning (English→target→English) with automatic 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, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Francisco Guzman, 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. In *nvm*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

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.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

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.

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

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.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *Preprint*, arXiv:1705.04304.

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.

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

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

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Preprint*, arXiv:2004.04696.

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.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and

20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Menno van Zaanen and Simon Zwarts. 2006. Unsupervised measurement of translation quality using multi-engine, bi-directional translation. In *AI 2006: Advances in Artificial Intelligence*, pages 1208–1214, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

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 Qualitative Examples

A.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.

A.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.

A.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!

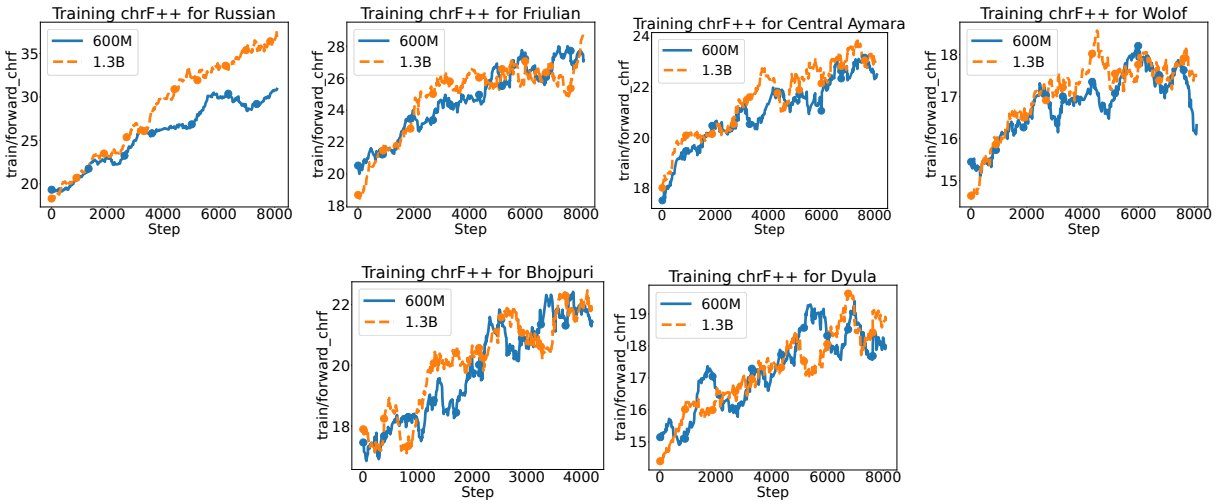

Figure 6: ChrF++ reward training curves after 8K optimization steps for the 6 languages for both model sizes (600M and 1.3B).

Skol'ko i kakie?
Trained (EN gloss): No problem! How many and what?

A.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
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 biiĠéej gu ñuul gi tasena ak Àll bu ñuul.; wala sax, B̈alkan folkloreḋaje 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 muy ànd 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 buum buum
Vanilla (EN gloss): In addition to its use as an anti-inflammatory, it is also used as an anti-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.

B Additional Training Plots

Figure 6 reports validation curves for the forward and backward translation models over 8K optimization steps.These curves are included for completeness and to illustrate the round-trip reinforcement learning training dynamics across languages. It's evident that there is no plateu and the models can further benefit from continued RL training.