Reference-Free Reinforcement Learning Fine-Tuning for MT: A Seq2Seq Perspective

Anonymous ACL submission

Abstract

Production machine translation relies overwhelmingly on encoder-decoder Seq2Seq models, yet reinforcement learning approaches to MT fine-tuning have largely targeted decoder-only LLMs at ≥7B parameters, with limited systematic study of encoder-decoder architectures. We apply Group Relative Policy Optimization to NLLB-200 (600M and 1.3B) using a hybrid reference-free reward — LaBSE and COMET-Kiwi — that requires no parallel data at fine-tuning time, evaluating across 13 typologically diverse languages. GRPO yields consistent improvements on all 13 languages, up to +5.03 chrF++ for Traditional Chinese, and, without any target-language data, competes with 3-epoch supervised fine-tuning on morphologically complex languages . We identify a consistent empirical pattern in which gains are largest where baseline performance is weakest and reward discriminability is highest, making this approach most effective precisely where parallel data is scarcest, and replicate this pattern across English and Spanish source languages.

1 Introduction

Encoder-decoder Seq2Seq models dominate production machine translation. They offer lower inference latency, smaller memory footprints, and stronger source-target alignment than autoregressive LLMs, making them the practical choice for deployment at scale, particularly for the long tail of language pairs where compute budgets are constrained (Costa-jussà et al., 2022). Yet the recent wave of reinforcement learning (RL) advances in NLP has almost entirely bypassed this architecture. Every application of Group Relative Policy Optimization (GRPO) to MT uses decoder-only LLMs of ≥7B parameters (He et al., 2025; Feng et al., 2025; Yang et al., 2025; Lu et al., 2025) — models that are impractical for production MT on most of the world's languages — with the sole exception of concurrent work by Attia and Fikri (2026), which we discuss below.

Two developments make this gap worth closing now. First, GRPO (Shao et al., 2024; Guo et al., 2025) has matured into a memory-efficient alternative to PPO that eliminates the value model, making RL fine-tuning accessible without specialized infrastructure. Second, reference-free quality estimators — LaBSE (Feng et al., 2022) and COMET-Kiwi (Rei et al., 2022) — have reached reliability levels that make them viable reward signals, enabling policy optimisation from monolingual source text alone. Together these developments create a practical path for RL-based MT improvement without parallel data, on the architectures practitioners actually deploy.

The key open question is not whether GRPO improves MT — it does, for high-resource language pairs on large decoder-only models. The question is what it costs to remove parallel supervision, and for which languages that cost is worth paying. Concurrent work by Attia and Fikri (2026) applies GRPO to NLLB-200, but uses English as the fixed source language and an indirect round-trip reconstruction objective, evaluating on six languages without SFT comparison or cross-domain analysis. No prior work has characterised how GRPO gains vary across typologically diverse languages spanning different scripts, morphological types, and baseline performance levels.

We present a systematic study of GRPO applied to NLLB-200 (600M and 1.3B) across 13 typologically diverse languages, trained on monolingual source text on a single NVIDIA A10G GPU. While the reward models (LaBSE and COMET-Kiwi) leverage parallel data during their own pre-training, our MT fine-tuning process remains reference-free as it requires only monolingual source text. Our contributions are:

- **Consistent reference-free gains.** GRPO im-

proves over the baseline on all 13 languages at both scales — up to +5.03 chrF++ for Traditional Chinese, competitive with 3-epoch SFT on morphologically complex languages without any target-language data, and transferring across domains on FLORES-200 and NTREX-128.

- **An empirical gain pattern.** Gain magnitude tends to be largest where baseline performance is weakest and reward discriminability is highest. This pattern replicates across English and Spanish source languages offering a potentially discriminative signal for practitioners selecting languages where reference-free RL is most likely to help.

- **Practical accessibility.** The full pipeline runs on a single 24 GB GPU with 4-bit quantization and LoRA, requires approximately 500 source sentences for reliable gains, and exhibits zero catastrophic forgetting on held-out languages.

2 Related Work

RL for sequence generation. Policy gradient methods for MT date back to Ranzato et al. (2016) and minimum risk training (Shen et al., 2016), both addressing exposure bias by directly optimising evaluation metrics. Ouyang et al. (2022) later established RLHF via PPO as the dominant alignment paradigm, at significant compute cost. GRPO (Shao et al., 2024; Guo et al., 2025) reduces this cost by eliminating the value model, normalising advantages from intra-group reward variation.

Group Relative Policy Optimization. Shao et al. (2024) introduced GRPO in DeepSeekMath as a memory-efficient alternative to PPO, eliminating the value model by normalising rewards within a group of K sampled outputs and estimating advantages from intra-group reward variation. Guo et al. (2025) subsequently scaled GRPO in DeepSeek-R1, demonstrating emergent reasoning capabilities and establishing it as the dominant post-training RL paradigm. A concurrent limitation of scalar reward models within GRPO was identified by Yang et al. (2026), who showed that evaluating hypotheses independently fails to discriminate fine-grained quality differences, and proposed a group-relative reward model (GRRM) that evaluates all K candidates jointly.

GRPO applied to MT. A growing body of work has applied GRPO to MT, all using large decoder-only LLMs. He et al. (2025) applied GRPO with a COMET reward to Qwen2.5-7B to induce chain-of-thought translation strategies. Feng et al. (2025) combined BLEU and COMET-Kiwi rewards within GRPO on 7B-parameter models, demonstrating emergent reasoning patterns and parity with proprietary systems on out-of-distribution tasks. Yang et al. (2025) introduced SSR-Zero, using the LLM itself as both generator and evaluator, with self-generated rewards combined with COMET signals achieving state-of-the-art performance. Lu et al. (2025) applied GRPO to Chinese-centric low-resource Southeast Asian MT via a semantic alignment reward and language-specific token prefixing. All of these works share three limitations our paper directly addresses: exclusive reliance on decoder-only LLMs of $\geq$7B parameters, focus on a narrow set of predominantly high-resource language pairs, and substantial compute requirements.

Reference-free quality estimation. COMET-Kiwi (Rei et al., 2022, 2023) enables quality estimation without target references by scoring source–hypothesis pairs against direct assessment annotations from professional translators. LaBSE (Feng et al., 2022) maps sentences into a shared cross-lingual embedding space, providing semantic similarity scores from source alone. Both were pre-trained on large parallel corpora; their reference-free property applies at inference. Kreutzer et al. (2017, 2018) examined reward shaping and feedback quality in RL-based MT, identifying the sensitivity of policy optimisation to reward noise.

Multilingual Seq2Seq MT. Our base model, NLLB-200 (Costa-jussà et al., 2022), is a 200-language encoder-decoder Seq2Seq model covering over 200 languages. Despite its breadth, NLLB-200 shows substantial performance variance across language families, with particularly weak scores on morphologically complex and low-resource languages. Koehn and Knowles (2017) identified morphological complexity as a core challenge for neural MT. Our work extends this to the RL setting by showing that morphological type and baseline performance jointly associate with GRPO gain magnitude in our study. In parallel, recent study by Attia and Fikri (2026) applies GRPO to NLLB-200 using a roundtrip method. Their approach uses English as the fixed source language and uses an indirect reconstruction objective rather than direct

quality estimation. Our work differs by using direct quality estimation on the forward translation, evaluating across 13 typologically diverse languages with an explicit SFT comparison and cross-domain analysis.

3 Methodology

3.1 GRPO for Seq2Seq MT

We frame MT as a RL problem in which the translation model serves as the policy π_θ, mapping a source sentence x to a target sentence y. At each step, the policy generates K candidate translations via temperature sampling. GRPO (Shao et al., 2024) estimates the advantage of each hypothesis directly from intra-group reward variation:

$$A_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r}) + \varepsilon} \tag{1}$$

where $\varepsilon = 10^{-4}$ is a stability floor that prevents noise amplification when reward variance collapses — a phenomenon we observe empirically in later training stages (§5). The policy is updated via a PPO-clipped surrogate objective (Schulman et al., 2017):

$$\mathcal{L}_{\text{clip}} = -\frac{1}{K}\sum_{i=1}^{K} \min(\rho_i A_i,\ \text{clip}(\rho_i,\ 1-\varepsilon_{\text{clip}},\ 1+\varepsilon_{\text{clip}}) \cdot A_i) \tag{2}$$

where $\rho_i = \exp(\log \pi_\theta(y_i|x) - \log \pi_{\text{ref}}(y_i|x))$ and $\varepsilon_{\text{clip}} = 0.2$. A KL penalty regularises against excessive deviation from the reference using the forward KL approximation from Shao et al. (2024):

$$\mathcal{L}_{\text{KL}} = E[\exp(\log \pi_\theta - \log \pi_{\text{ref}}) - (\log \pi_\theta - \log \pi_{\text{ref}}) - 1] \tag{3}$$

The full objective is $\mathcal{L} = \mathcal{L}_{\text{clip}} + \beta \cdot \mathcal{L}_{\text{KL}}$, where β is investigated in §4.5. A key implementation detail for Seq2Seq models is that π_{ref} is obtained by disabling the LoRA adapters on the same model — no separate reference model is required. This anchors the policy to NLLB's pretrained multilingual knowledge and prevents catastrophic forgetting.

3.2 Hybrid Reference-Free Reward

Both reward components are reference-free at fine-tuning time but were pre-trained on parallel corpora; the reference-free property applies to the fine-tuning stage only. The first component, LaBSE (Feng et al., 2022), provides cross-lingual semantic adequacy via cosine similarity between normalised source and hypothesis embeddings. The second, COMET-Kiwi (Rei et al., 2022, 2023), provides a learned quality estimate from direct assessment annotations by professional translators. The hybrid reward combines both with equal weighting:

$$r_{\text{hyb}}(x, y_i) = \frac{1}{2}\big(r_{\text{LaBSE}}(x, y_i) + r_{\text{COMET}}(x, y_i)\big) \tag{4}$$

LaBSE guards against reward hacking toward fluent but unfaithful translations; COMET-Kiwi contributes a richer signal aligned with human quality judgements. Equal weighting is adopted as a language-agnostic default and validated in the discriminability analysis below.

3.3 Reward Discriminability Analysis

To validate that the hybrid reward produces meaningful quality rankings prior to training, we evaluate it on a quality cline of six candidate translations per source sentence (50 sentences per language from FLORES-200 dev), computing Pearson r between reward score and quality rank across five LaBSE/COMET-Kiwi weight configurations. All configurations achieve good-to-excellent discrimination (mean r ranging from -0.90 to -0.94), demonstrating robustness to the exact weighting. LaBSE-only is optimal for Yoruba ($r = -0.94$ vs. -0.81 for COMET-Kiwi alone), reflecting limited COMET-Kiwi calibration on underrepresented languages; the hybrid achieves highest discrimination for morphologically rich languages (Arabic, Belarusian). Table 11 (Appendix I) reports full results. We adopt 0.50/0.50 equal weighting as the language-agnostic default, supported by an analysis of the two components' complementarity: across 5,060 baseline translation hypotheses spanning four of the studied languages (Basque, Bengali, Yoruba, and Traditional Chinese), LaBSE and COMET-Kiwi scores show only moderate correlation (Pearson $r = 0.528$), confirming that the two signals are not redundant and that their combination captures quality dimensions neither component covers alone.

3.4 Base Model and Parameter-Efficient Fine-Tuning

We use NLLB-200 (No Language Left Behind; Costa-jussà et al., 2022) as our base model, evaluating both the distilled 600M and 1.3B parameter variants. To enable training on a single GPU, we quantise model weights to 4-bit NF4 precision (Dettmers et al., 2023) and apply LoRA (Hu et al., 2022) to the query and value projections of all attention layers (rank r=16, α=32, dropout 0.05). Only the LoRA adapter weights are updated; base

model weights remain frozen. Peak VRAM usage is approximately 4 GB (600M) and 6 GB (1.3B) on a single NVIDIA A10G (24 GB). Full hyperparameters are in Appendix B.

3.5 Experimental Design

We conduct two main experiments, followed by controlled ablations. Our 13 languages span 9 language families, 7 scripts, and 4 morphological types (Appendix A). In both experiments, NLLB-200 translates from English (`eng_Latn`) as the fixed source language; all reward scoring is over English→target pairs.

Experiment A: Systematic Multiple Target Evaluation. Training uses the FLORES-200 development split (Goyal et al., 2022) (≈997 sentences per language), source side only, for 3 epochs (≈3,000 steps). We evaluate on FLORES-200 devtest (1,012 sentences) and compare three systems: the unfine-tuned NLLB baseline, an SFT adapter with identical LoRA configuration trained on parallel FLORES-200 data, and the GRPO adapter. Both model variants are evaluated.

Experiment B: Cross-Domain Adaptation. Training uses 10,000 English sentences from CC-News with no target-language text at any stage, for 10,000 steps. We evaluate on FLORES-200 devtest (out-of-domain) and NTREX-128 (Federmann et al., 2022) (news domain, 1,997 sentences), testing both out-of-domain and closer-domain transfer.

3.6 Ablation Studies

We conduct two ablations using Basque (`eus_Latn`) and Traditional Chinese (`zho_Hant`) on the 600M model. The **KL regularisation** ablation varies $\beta \in \{0.0, 0.001, 0.01, 0.05\}$ to test whether the quality ceiling reflects regularisation strength or reward discriminability. The **training data size** ablation varies $N \in \{100, 250, 500, 1{,}000\}$ to characterise the minimum data requirement and GRPO scaling behaviour relative to SFT.

3.7 Evaluation Metrics

Our primary metric is chrF++ (Popović, 2015), a character n-gram F-score well-suited to morphologically complex languages. BLEU (Papineni et al., 2002) is reported via SacreBLEU (Post, 2018) for comparability with prior work. For fully independent neural validation we report COMET-22 (reference-based, no overlap with the reward family) and BERTScore F1 (Zhang et al., 2019) using `bert-base-multilingual-cased` with baseline rescaling, computed against human references independently of both the reward signal and the chrF++ family. chrF++ is our primary metric for all comparisons; it is entirely independent of the reward signal.

4 Results

4.1 Experiment A: Systematic Multiple Target Performance

Across 13 languages and 2 model scales evaluated in Experiment A, reference-free GRPO consistently improves over the NLLB baseline. Gains are robust across diverse scripts — Cyrillic, Brahmic, Hanzi, Tibetan — and grammatical structures, demonstrating that the hybrid reward signal generalises well beyond the high-resource language pairs it was calibrated on.

A consistent gain pattern. The largest gains occur in Traditional Chinese (+5.03 chrF++), which has the weakest baseline in our set (14.47 chrF++), partly reflecting documented confusion in NLLB-200's internal representations between Traditional and Simplified Chinese (Caswell et al., 2023). Tibetan (+3.01, 600M) and Basque (+2.57, 1.3B) follow the same pattern: weak baselines and reward discriminability scores above the dataset mean (Table 11). Conversely, Arabic and Belarusian show the smallest improvements despite high reward discriminability scores, confirming that discriminability is a necessary but not sufficient condition for large gains. Swahili appears to contradict this pattern — its baseline (60.33 chrF++) is the strongest in our set yet it gains +2.30 — but its discriminability score (0.945) is well above the dataset mean, consistent with headroom magnitude where high discriminability can partially offset a strong baseline. This suggests the two factors may interact rather than operate independently.

Morphological type and gain magnitude. Fusional languages show the smallest average gains (mean +0.82 chrF++, 1.3B), while agglutinative (+1.71) and isolating (+2.09) languages improve more. However, morphological type is a weak predictor at best: Japanese (agglutinative) gains only +0.54 chrF++ while Traditional Chinese (isolating) gains +5.03, and within each morphological group baseline performance explains more variance than

typological class. We therefore treat morphological type as a coarse organizing lens rather than a causal factor, with baseline performance the more consistent factor of gain magnitude in this study.

A frugal human preference study on four languages confirms that large automatic gains are human-perceptible while small gains are not, consistent with established perceptibility thresholds in MT evaluation (Appendix K).

4.2 GRPO versus Supervised Fine-Tuning

A natural question for any reference-free RL approach is how it compares to SFT given access to the same data. We do not frame this as a competition — SFT and GRPO operate under fundamentally different assumptions, SFT requiring parallel supervision and GRPO requiring none — but rather as a calibration exercise: understanding where each method is strong, where each struggles, and what the cost of removing parallel data actually is. Importantly, the SFT baseline trains on the parallel version of the same FLORES-200 development sentences used as source-only input for GRPO, meaning SFT has seen the target side of the evaluation-domain data during training while GRPO has not. GRPO's competitiveness under this asymmetry — not SFT's inferiority as a method — is the meaningful finding. Results are shown in Table 1.

At comparable compute, GRPO and SFT are close. Against the 1-epoch SFT baseline, the two methods produce similar chrF++ scores on most languages. GRPO has an edge on morphologically complex languages with weak baselines —Basque (+1.76 chrF++ over $\text{SFT}_{1\text{ep}}$), Bengali (+0.87), Swahili (+0.75) — while SFT is stronger on Arabic (+0.43 vs. +0.10) and Yoruba. The key asymmetry is the supervision cost: GRPO achieves these results *without a single target-language sentence*, while SFT requires fully aligned parallel data for every language.

Three-epoch SFT is stronger on average, but the gap is modest. Given three full epochs of parallel data, SFT outperforms GRPO on Swahili, Turkish, Japanese, Yoruba, and both Chinese variants. However, the margins are generally small: for Czech, Polish, Bengali, and Belarusian, the difference is within ±0.2 chrF++ and non-significant, meaning GRPO with no parallel data matches three epochs of supervised training on these languages. GRPO retains a clear advantage on Basque (+0.51 chrF++ over $\text{SFT}_{3\text{ep}}$), the language with the weakest NLLB baseline among the fusional and agglutinative groups. The overall picture is one of competitive parity on different contexts rather than decisive superiority for either method.

Arabic is the clearest SFT-favoured case. Arabic is the only language where SFT consistently outperforms GRPO across both epoch conditions, and where the gap is practically meaningful: SFT yields +1.27 chrF++ at 3 epochs while GRPO yields +0.10. This is not a failure of the training procedure but a reflection of Arabic's operating conditions — strong baseline (54.72 chrF++), limited headroom, and weaker COMET-Kiwi calibration for root-pattern morphology (§5) — which make supervised signal more effective than reward-guided exploration for this language.

Independent metrics reveal complementary strengths. COMET-22 and BERTScore tell a more nuanced story than chrF++ alone. For Japanese, GRPO yields +0.018 COMET-22 versus SFT's +0.006 — a threefold difference despite nearly identical chrF++ scores — and BERTScore ranks Japanese third among all languages (+0.012 F1). This suggests GRPO finds translation paths with better semantic adequacy than a single parallel reference recovers, an effect chrF++ partially obscures. For Yoruba, SFT degrades COMET-22 by −0.022 while GRPO gains +0.004, illustrating a practical risk of supervised fine-tuning on limited parallel data for underrepresented languages: it can hurt neural quality alignment even when surface metrics improve. These divergences suggest the two methods are not simply trading off on the same quality dimension, but exploring different regions of the translation space.

4.3 Effect of Model Scale

The 1.3B model achieves larger gains on 7 of 13 languages, with the most pronounced scaling effect on Traditional Chinese (+3.94 vs. +5.03 chrF++) and Turkish (+0.91 vs. +1.80). Five languages show reversed scaling: Tibetan (+3.01 vs. +1.33), Bengali (+1.98 vs. +1.13), Yoruba (+0.93 vs. +0.12), Simplified Chinese (+1.73 vs. +1.12), and Belarusian (+0.36 vs. +0.14). For Tibetan, the reversal is corroborated by COMET-22 (+0.004 vs. −0.002); for Yoruba, the 1.3B model's negative COMET-22 delta (−0.006) suggests reward hacking rather than genuine improvement. We hypothesise that for low-resource languages with noisier

			ΔchrF++			ΔC-22			ΔBS		
Group	**Lang.**	**Base**	**SFT_{1ep}**	**SFT_{3ep}**	**GRPO**	**SFT_{1ep}**	**SFT_{3ep}**	**GRPO**	**SFT_{1ep}**	**SFT_{3ep}**	**GRPO**
Aggl.	eus	52.69	+1.10	+2.48	**+2.93**	+0.005	+0.011	**+0.012**	+0.003	+0.005	**+0.006**
	swh	60.33	+1.18	**+2.70**	+1.88	+0.003	**+0.008**	+0.004	+0.005	**+0.011**	+0.008
	bod	23.37	+1.43	+2.85	**+3.01**	+0.035	**+0.068**	+0.004	+0.017	**+0.032**	+0.012
	tur	56.49	+0.68	**+1.80**	+0.95	+0.002	**+0.005**	+0.001	+0.002	**+0.005**	+0.002
	jpn	29.31	+0.96	**+1.46**	+0.86	+0.012	+0.006	**+0.018**	+0.009	+0.007	**+0.011**
Isol.	zho_Hant	16.60	+3.74	**+5.67**	+5.07	+0.069	**+0.100**	+0.092	+0.102	**+0.138**	+0.115
	zho_Hans	24.59	+2.29	**+2.88**	+2.28	+0.030	**+0.039**	+0.034	+0.062	**+0.071**	+0.040
	yor	24.28	+1.28	**+1.56**	+0.97	−0.011	−0.022	**+0.004**	+0.000	−0.001	**+0.007**
Root-pat.	arb	54.72	+0.41	**+1.27**	+0.10	+0.004	**+0.006**	−0.000	+0.002	**+0.005**	+0.001
Fus.	bel	41.75	+0.10	+0.43	**+0.55**	+0.003	+0.007	−0.008	+0.000	+0.000	+0.000
	ben	49.84	+1.07	+2.15	**+1.99**	+0.003	**+0.007**	+0.004	+0.003	**+0.007**	+0.006
	ces	55.13	+0.17	**+1.08**	+0.57	+0.002	**+0.005**	+0.001	+0.001	**+0.003**	+0.002
	pol	48.31	+0.27	**+0.68**	+0.55	+0.002	**+0.003**	−0.000	+0.001	+0.001	**+0.001**

Table 1: Comparison of GRPO and SFT on FLORES-200 devtest (600M model). All columns report deltas over the unfine-tuned baseline. Underlined values indicate statistical significance ($p < 0.05$, paired bootstrap resampling, n=1,000). **Bold** denotes the highest value per row within each metric block. ΔC-22: COMET-22 delta (reference-based; shares architectural lineage with COMET-Kiwi; chrF++ remains the primary metric). ΔBS: BERTScore F1 delta (`bert-base-multilingual-cased`, unrescaled) against human references. GRPO outperforms SFT_{1ep} by BERTScore on 10 of 13 languages (exceptions: Arabic, Belarusian, Czech where gains are within ±0.001).

reward signals, the 600M model's lower capacity acts as an implicit regulariser. Averaged across languages where the 1.3B is stronger, the 600M recovers roughly 80% of its gains at 4 GB peak VRAM versus 6 GB, making it the practical default for low-resource deployment.

4.4 Experiment B: Cross-Domain Adaptation

Experiment B trains on 10,000 monolingual English sentences from CCNews with no target-language supervision and evaluates on FLORES-200 devtest (out-of-domain) and NTREX-128 (closer-domain news). Full results are in Appendices D and E.

Cross-domain gains largely replicate Experiment A. This pattern holds across domains: Traditional Chinese (1.3B: +4.88 chrF++), Swahili (+3.00), and Basque (+2.57) show the strongest FLORES-200 gains, confirming that the reward signal generalises beyond the in-domain training distribution. Arabic is the only language degrading at the best checkpoint (600M: −0.13 chrF++), and Tibetan requires best-checkpoint selection — without early stopping, reward variance collapse reduces chrF++ below 2.0 under the 10,000-step schedule (§5).

NTREX-128: domain proximity favours GRPO on low-resource scripts. On the news-domain NTREX benchmark — closer to CCNews than to FLORES-200 — GRPO outperforms 3-epoch SFT on four languages: Basque, Tibetan, Japanese, and Traditional Chinese (600M: + 3.74 vs. SFT +2.03; 1.3B: +5.22). SFT retains an advantage on Arabic, Yoruba, Polish, Czech, and Bengali, where parallel supervision is more effective than reward-guided exploration at competitive baselines. Belarusian degrades below baseline for both systems (−0.63 GRPO, −0.54 SFT), indicating domain mismatch independent of the training objective. Overall, GRPO is the more practical option where parallel data is unavailable and baseline headroom is large; SFT remains preferable where calibrated supervision is available.

4.5 Ablation Studies

Two ablations use Basque (`eus_Latn`) and Traditional Chinese (`zho_Hant`) on the 600M model.

KL regularisation. Table 2 reports best chrF++ across $\beta \in \{0.0, 0.001, 0.01, 0.05\}$. Results are insensitive to this choice: the spread is 0.19 chrF++ for Basque and 0.61 for Chinese across all four settings, and $\beta = 0.0$ performs on par with regularised variants. This confirms that the LoRA adapter provides sufficient structural constraint against policy drift without an explicit KL term, and that the quality ceiling reflects the intrinsic discriminability limit of the hybrid reward rather than excessive regularisation (§5).

Training data size. Table 3 reports chrF++ across $N \in \{100, 250, 500, 1{,}000\}$ sentences. A

Lang.	$\beta = 0.0$	$\beta = 0.001$	$\beta = 0.01$	$\beta = 0.05$
eus	51.14	51.03	**51.23**	51.14
zho_Hant	**19.20**	18.59	18.97	18.77

Table 2: Best chrF++ across KL regularisation coefficients β (600M). Spread of 0.19 (Basque) and 0.61 (Chinese) across all settings confirms robustness to β choice.

reliability threshold emerges at $N = 500$: below this, neither SFT nor GRPO produces significant gains. At $N = 500$ both methods reach significance simultaneously with nearly identical scores. At $N = 1{,}000$ they diverge: GRPO significantly outperforms SFT for Basque (+0.26 chrF++, $p = 0.019$), where source diversity generates sufficient intra-group reward variance; for Traditional Chinese the difference remains non-significant ($p = 0.304$), consistent with the reward signal approaching its discriminability ceiling. Further analysis in the Appendix G.

	eus (Base: 47.64)		zho_Hant (Base: 14.12)	
N	**SFT**	**GRPO**	**SFT**	**GRPO**
100	47.78	47.75	14.24	14.11
250	47.79	47.74	14.11	14.17
500	48.12	**48.17**	14.80	**14.81**
1000	48.74	**48.99**	**16.86**	16.79

Table 3: chrF++ across training set sizes N (600M model). **Bold** denotes the higher of SFT and GRPO at each N. Neither method produces statistically significant gains below $N = 500$; at $N = 1{,}000$, GRPO significantly outperforms SFT for Basque ($p = 0.019$) but not for Traditional Chinese, where both methods approach the reward discriminability ceiling.

5 Analysis

5.1 Empirical Predictors of Gain Magnitude

Table 4 reports baseline chrF++, reward discriminability, and GRPO ΔchrF++ for 13 languages. Across the full sample ($n = 13$), the Spearman correlation between baseline chrF++ and ΔchrF++ is low ($\rho = 0.28$, $p = 0.35$). Traditional Chinese warrants separate treatment: Caswell et al. (2023) documented that NLLB-200 cannot reliably distinguish it from Simplified Chinese, producing script-mixed outputs — a known model defect independent of RL dynamics. Controlling for this artefact ($n = 12$), the correlation strengthens to $\rho = 0.67$ ($p = 0.018$). This estimate is robust: a 10,000-resample bootstrap yields 95% CI $[0.17, 0.98]$, and leave-one-out analysis preserves significance across all 12 folds ($p < 0.05$).

To assess whether the pattern reflects target-language properties rather than English-specific training dynamics, we run GRPO with Spanish (spa_Latn) source sentences on 7 languages under identical hyperparameters. The rank ordering of gains is highly consistent with English-source results (Spearman $\rho = 0.893$, $p = 0.007$, $n = 7$): Traditional Chinese remains the largest gainer (+4.05 chrF++ from Spanish vs. +3.94 from English), Arabic remains the failure case (+0.12 vs. +0.08), and Tibetan retains a large gain (+2.71 vs. +3.01). This cross-source consistency provides independent replication of the pattern under a distinct experimental condition, confirming that gain magnitude is driven by target-language properties rather than source-language training dynamics. Note that pooling Spanish and English observations is not meaningful since absolute baseline values are source-language dependent — Spanish baselines are uniformly lower than English baselines for the same languages. Full results and per-language discussion are in Appendix F.

We present this as an exploratory empirical regularity that may inform practitioner decisions: gains are largest where baseline quality is lowest and reward signals are clearest. BERTScore F1 gains correlate strongly with chrF++ gains ($r = 0.81$, $p < 0.001$), and Experiments A and B BERTScore deltas correlate at $r = 0.998$ ($p < 0.001$), confirming consistency across independent metric families and training domains. Establishing this as a general principle requires validation on a larger language sample, segment-level analysis, and reward ablations decoupling LaBSE and COMET-Kiwi contributions per language family.

5.2 Failure Modes

Arabic. Arabic consistently fails to benefit from reference-free GRPO across all conditions, due to the convergence of three factors: a strong baseline (51–54 chrF++) leaving limited headroom; root-pattern morphology that COMET-Kiwi — trained predominantly on Indo-European and Sino-Tibetan pairs — cannot reliably score; and domain mismatch under CCNews training. Together these place Arabic outside the current operating regime of the hybrid LaBSE–COMET-Kiwi reward.

Tibetan and reward variance collapse. Tibetan shows healthy gains at the best checkpoint (600M: +3.01 chrF++) but collapses to below 2.0 chrF++ in

Lang.	Morph.	Baseline	Disc.	ΔchrF++
zho_Hant	Isol.	14.5	0.885	+5.03†
zho_Hans	Isol.	21.8	0.887	+1.12
yor	Isol.	24.5	0.929	+0.12
jpn	Aggl.	25.5	0.864	+0.54
bod	Aggl.	26.5	0.872	+1.33
bel	Fus.	40.6	0.958	+0.14
pol	Fus.	47.2	0.949	+0.70
ben	Fus.	47.8	0.943	+1.13
eus	Aggl.	49.9	0.947	+2.57
arb	Root.	54.4	0.964	+0.68
ces	Fus.	55.2	0.955	+1.29
tur	Aggl.	56.1	0.940	+1.80
swh	Aggl.	59.0	0.945	+2.30

Spearman ρ (baseline vs. ΔchrF++):
Full (n=13): $\rho = 0.28$, $p = 0.35$
Excl. † (n=12): $\rho = 0.67$, $p = 0.018^*$

Table 4: Baseline chrF++, reward discriminability (hybrid 0.50/0.50), and GRPO ΔchrF++ (1.3B, Experiment A), sorted by baseline. † Traditional Chinese outlier: low baseline compounded by NLLB-200 script-confusion (Caswell et al., 2023). (Table 11).

Experiment B without early stopping. As training progresses, all K hypotheses converge to similar reward scores, driving $\text{std}(\mathbf{r}) \to \varepsilon$ in Equation 1. The resulting numerical instability amplifies small reward differences into large parameter updates that push the policy into degenerate output modes. The $\varepsilon = 10^{-4}$ floor is sufficient for Experiment A's 3,000-step schedule but fails at 10,000 steps. Best checkpoint selection on a held-out subset is a necessary safeguard for extended training.

Human preference evaluation. To assess whether automatic metric gains reflect human-perceptible quality improvements, we conduct a preference study across four languages spanning a range of GRPO gain magnitudes. Annotators evaluated 50 sentence pairs per language (GRPO vs. Baseline, blinded) and indicated a preference or tie. Results reveal a clear *perceptibility threshold*: for Traditional Chinese, where GRPO achieves +3.94 chrF++ (600M), annotators strongly prefer GRPO (68% vs. 8% Baseline, $p < 0.001$). For Turkish (+0.91 chrF++), Bengali (+1.98), and Polish (+0.41), preferences are not significantly different from chance (all $p > 0.5$). This pattern is consistent with established findings in MT human evaluation that sub-2-point chrF++ improvements are typically sub-perceptual (Graham et al., 2015), and suggests that GRPO's largest gains — concentrated in languages with the weakest baselines — are the ones most likely to be noticed by users. The absence of significant preference for smaller gains should not be interpreted as evidence of reward hacking; Turkish and Polish show exact or near-exact parity rather than Baseline preference, indicating the policy is not degrading translation quality even when gains are small.

5.3 Catastrophic Forgetting

Across 39 GRPO adapter evaluations on three held-out languages (French, Hindi, Russian), **zero** forgetting events are observed (mean delta +0.34 chrF++, range −0.16 to +1.13). The single forgetting event in the full dataset is an SFT adapter: Belarusian SFT degrades Hindi by −0.75 chrF++. Freezing the base model weights and updating only LoRA adapters preserves NLLB-200's pretrained cross-lingual representations throughout training. Full results are in Appendix H.

6 Conclusion

We presented a systematic study of GRPO applied to a Seq2Seq MT model across 13 typologically diverse languages using a hybrid reference-free reward requiring no parallel data at fine-tuning time. GRPO yields consistent gains across scripts and morphological types, reproducible on a single GPU, and corroborated by three independent validators (COMET-22, BERTScore F1, and cross-domain stability). The central observation is a consistent empirical pattern whereby gains tend to be largest where baselines are weakest and reward discriminability is highest — an association that warrants broader validation but offers a potentially useful signal for practitioners. Against SFT, GRPO is competitive at one epoch and matches three epochs on morphologically complex languages without any target-language data. Arabic and Tibetan identify the current boundaries: the former resists improvement due to strong baseline and reward calibration gaps; the latter requires early stopping to avoid reward variance collapse. Both failure modes point to the same underlying constraint — the quality ceiling is set by reward discriminability, not by the policy optimisation. Future work should explore morphology-aware reward components, adaptive variance regularisation, and multilingual joint GRPO training.

Limitations

Reward pre-training on parallel data. Although no parallel data is required at fine-tuning time, both reward components — LaBSE and

COMET-Kiwi — were themselves pre-trained on large parallel corpora. The reference-free property therefore applies to the fine-tuning stage only. For languages entirely absent from both reward models' training distributions, the quality signal may be unreliable, limiting the approach's applicability to genuinely zero-resource settings where the target language has no representation in existing multilingual encoders.

Reward discriminability for low-resource languages. Our discriminability analysis uses a constructed quality cline rather than actual GRPO rollouts. Intra-group reward variance on real model outputs would constitute a stronger validation of the reward signal's effectiveness during training. Furthermore, COMET-Kiwi calibration degrades for languages underrepresented in its training data, as evidenced by Arabic's consistent non-improvement and Yoruba's reversed scaling between model sizes. Extending the hybrid reward with language-family-specific components would likely improve coverage for these cases.

English-only source language. All experiments use English (`eng_Latn`) as the fixed source language. Whether the findings generalise to non-English source languages, where reward model calibration for source-hypothesis pairs may differ, remains an open question. The scope of this work refers to the diversity of target languages, not source languages. We provide a cross-source replication study with Spanish source sentences (Appendix F), which confirms the gain pattern holds across source languages.

Scale and generalisability. All experiments use NLLB-200 at 600M and 1.3B parameters. Whether the observed dynamics, the empirical gain pattern, reversed scaling, reward variance collapse, generalise to other encoder-decoder architectures, larger models, or languages beyond our 13-language set remains an open question. The 13-language sample, while typologically diverse, is insufficient to establish statistically significant correlations between linguistic properties and GRPO gain magnitude, and our findings should be treated as hypotheses motivating further investigation rather than universal laws.

Training instability under extended schedules. Tibetan's catastrophic collapse in Experiment B demonstrates that GRPO is susceptible to reward variance collapse under long training schedules, and that the $\varepsilon = 10^{-4}$ stability floor is insufficient to prevent this in all conditions. Best-checkpoint selection mitigates the problem but does not eliminate it. Practitioners applying GRPO to new languages or domains should treat early stopping on a held-out development set as a required component of the training pipeline rather than an optional refinement.

References

Ahmed Attia and Alham Fikri. 2026. Improving low-resource machine translation via round-trip reinforcement learning. *arXiv preprint arXiv:2601.12535*.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tim Dettmers, Artidoro Pagnoni, Ari Rodola, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Speech and Language Research*, pages 21–28.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 878–891.

Zhaopeng Feng, Ruidi Cai, Jiaxuan Liu, Junyuan Hu, and Zhiyong Wu. 2025. MT-R1-Zero: Advancing LLM-based machine translation via R1-Zero-style reinforcement learning. *arXiv preprint arXiv:2503.06973*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-200 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics (TACL)*, 10:522–538.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of NAACL*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxian Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Minggui He, Zhiwei Li, Shanshan Li, Hang Peng, Shimin Zhao, Yuang Li, Jiaxin Luo, Chang Hao, Shiyue Guo, Rui Li, and 1 others. 2025. R1-T1: Fully incentivizing translation capability in LLMs via reasoning learning. *arXiv preprint arXiv:2502.19735*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation (WNMT)*, pages 28–39.

Julia Kreutzer, Artem Sokolov, and Stefan Riezler. 2017. Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1503–1513.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Quality estimation from scratch (QUETCH): Deep learning approaches for multilingual word-level translation quality estimation. In *Proceedings of the Third Conference on Machine Translation (WMT)*, pages 801–810.

Wenhao Lu, Xuebo Wang, Min Zhang, and Runzhe Zhan. 2025. MERIT: Multilingual semantic alignment reward for machine translation via reinforcement learning. *arXiv preprint arXiv:2504.01496*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744.

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 (ACL)*, pages 311–318.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*, pages 186–191.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-Kiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Chrysoula Zerva, Ana C. Farinha, Duarte Maroti, José G. C. de Souza, Luisa Coheur, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 863–878.

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

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

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1683–1692.

Sen Yang, Shanbo Cheng, Lu Xu, Jianbing Zhang, and Shujian Huang. 2026. GRRM: Group relative reward modeling for machine translation. *Preprint*, arXiv:2602.14028.

Yu Yang, Shanbo Cheng, Lu Xu, Jianbing Zhang, and Shujian Huang. 2025. SSR-Zero: Simple self-rewarding reinforcement learning for machine translation. *arXiv preprint arXiv:2503.16681*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Language Set

Table 5 provides the complete set of 13 languages used across all experiments, including FLORES-200/NLLB language codes, language family, script, and morphological type.

Code	Language	Family	Script	Morphology
eus_Latn	Basque	Isolate	Latin	Agglutinative
tur_Latn	Turkish	Turkic	Latin	Agglutinative
swh_Latn	Swahili	Bantu	Latin	Agglutinative
pol_Latn	Polish	Slavic	Latin	Fusional
ces_Latn	Czech	Slavic	Latin	Fusional
ben_Beng	Bengali	Indo-Aryan	Bengali	Fusional
arb_Arab	Arabic	Semitic	Arabic	Root-pattern
bel_Cyrl	Belarusian	Slavic	Cyrillic	Fusional
bod_Tibt	Tibetan	Sino-Tibetan	Tibetan	Agglutinative
yor_Latn	Yoruba	Niger-Congo	Latin	Isolating
jpn_Jpan	Japanese	Japonic	Japanese	Agglutinative
zho_Hant	Chinese (T)	Sino-Tibetan	Traditional	Isolating
zho_Hans	Chinese (S)	Sino-Tibetan	Simplified	Isolating

Table 5: Complete language set. Morphological types: agglutinative (6), fusional (4), isolating (3), root-pattern (1). Script families: Latin (6), CJK (2), Japanese (1), Tibetan (1), Arabic (1), Bengali (1), Cyrillic (1).

B Hyperparameters

Table 6 lists all fixed hyperparameters used across Experiments A and B. The only hyperparameter varied across runs is the KL regularisation coefficient β, studied in the ablation of §4.5.

Hyperparameter	Value
Base models	NLLB-200-distilled-1.3B / 600M
LoRA rank r	16
LoRA α	32
LoRA target modules	q_proj, v_proj
LoRA dropout	0.05
Quantization	4-bit NF4, bfloat16
Hypotheses per step K	12
KL coefficient β	0.001 (except ablation)
PPO clip ε	0.2
Learning rate	5×10^{-6}
Optimizer	AdamW
Max tokens (train)	64
Max tokens (eval)	128
Temperature	1.2
Eval frequency	every 50 steps
Eval subset size	100 sentences
Hardware	Single NVIDIA A10G (24 GB)

Table 6: Fixed hyperparameters used across all experiments.

C Full Experiment A Results

Table 7 reports complete Experiment A results including absolute chrF++ and BLEU scores, COMET-22 deltas, and BERTScore F1 deltas for both the 600M and 1.3B model variants. The condensed version reporting only ΔchrF++ and ΔC-22 appears in the main body as Table 1.

BLEU scores for Traditional Chinese (zho_Hant), Simplified Chinese (zho_Hans), Japanese (jpn_Jpan), and Tibetan (bod_Tibt) should be interpreted with caution: word-level tokenisation is poorly calibrated for these scripts, causing BLEU to diverge from chrF++, COMET-22, and BERTScore. We report BLEU for completeness and comparability with prior work only.

D Experiment B: Full FLORES-200 Results

Table 8 reports complete Experiment B results on FLORES-200 devtest for both model sizes at the best-performing checkpoint. For Tibetan (bod_Tibt), best-checkpoint selection is essential: while the best checkpoint yields healthy gains (600M: +2.47, 1.3B: +0.68 chrF++), training without early stopping causes catastrophic output collapse under the 10,000-step CCNews schedule, reducing chrF++ to below 2.0 for both model sizes. This behaviour is discussed in §5.

E NTREX-128 Results

Table 9 reports chrF++ on the NTREX-128 news-domain benchmark for the 600M and 1.3B models. GRPO is trained on CCNews monolingual source text (Experiment B); the $\text{SFT}_{3\text{ep}}$ baseline is trained on FLORES-200 parallel data and is only available for the 600M model. NTREX serves as a closer-domain evaluation for GRPO and a cross-domain evaluation for SFT. Key findings are summarised in §4.4.

F Cross-Source Replication: Spanish Source Language

To assess whether the empirical gain pattern reflects target-language properties rather than English-specific training dynamics, we run GRPO with Spanish (spa_Latn) source sentences on 7 languages under identical hyperparameters (600M

Group	Lang.	Size	Baseline	ΔchrF++	ΔBLEU	ΔC-22	ΔBS
Aggl.	eus	600M	47.75	+2.44	+0.15	+0.012	+0.006
		1.3B	49.88	+2.57	+0.87	+0.010	+0.006
	swh	600M	57.79	+1.94	+2.07	+0.004	+0.008
		1.3B	58.96	+2.30	+2.46	+0.008	+0.009
	bod	600M	23.37	+3.01	+0.70	+0.004	+0.006
		1.3B	26.45	+1.33	+0.09	−0.002	+0.003
	tur	600M	52.68	+0.91	+0.81	+0.001	+0.002
		1.3B	56.07	+1.80	+1.58	+0.003	+0.006
	jpn	600M	23.45	+0.49	+1.56†	+0.018	+0.011
		1.3B	25.54	+0.54	+1.40†	+0.015	+0.012
Isol.	zho_Hant	600M	13.90	+3.94	+8.12†	+0.092	+0.115
		1.3B	14.47	**+5.03**	**+9.20†**	**+0.086**	**+0.115**
	zho_Hans	600M	19.22	+1.73	+3.87†	+0.034	+0.040
		1.3B	21.83	+1.12	+3.10†	+0.027	+0.036
	yor	600M	22.70	+0.93	+0.33	+0.004	+0.007
		1.3B	24.52	+0.12	+0.08	−0.006	+0.003
Root-pat.	arb	600M	51.12	+0.08	−0.20	−0.000	+0.001
		1.3B	54.38	+0.68	+0.58	+0.003	+0.004
Fus.	bel	600M	38.20	+0.36	−0.34	−0.008	+0.000
		1.3B	40.57	+0.14	+0.04	+0.000	+0.000
	ben	600M	45.56	+1.98	+1.64	+0.004	+0.006
		1.3B	47.80	+1.13	+0.85	+0.001	+0.003
	ces	600M	52.54	+0.50	+0.10	+0.001	+0.004
		1.3B	55.18	+1.29	+0.72	+0.004	+0.012
	pol	600M	45.01	+0.41	−0.20	−0.000	+0.001
		1.3B	47.16	+0.70	+0.02	+0.002	+0.002

Table 7: Complete Experiment A results (FLORES-200 devtest, *grpo_best* checkpoint) for both model sizes. **Baseline** reports absolute chrF++ for the unfine-tuned NLLB model; all other columns report deltas over the baseline. Underlined values indicate statistical significance ($p < 0.05$, paired bootstrap resampling, n=1,000). ΔC-22 is the independent reference-based COMET-22 validator. ΔBS is the BERTScore F1 delta (`bert-base-multilingual-cased`, unrescaled) against human references, independent of both the reward and chrF++ metric families. **Bold** denotes the largest absolute gains per metric. † ΔBLEU for Chinese uses the `zh` character-level tokenizer and for Japanese the `char` tokenizer; previously reported negative values reflected incorrect whitespace tokenization. Tibetan (`bod`) BLEU is retained for completeness but remains unreliable due to the absence of a standard tokenizer.

model, Experiment A schedule, 3 epochs). We evaluate on FLORES-200 devtest using Spanish source sentences throughout; all reward scoring is over Spanish→target pairs. SFT baselines are not included for this condition as our primary comparison focuses on whether the relative ordering of GRPO gains is preserved across source languages.

Table 10 reports results alongside the corresponding English-source gains from Experiment A. Spanish baselines are systematically lower than English baselines across all 7 languages (mean −5.6 chrF++), reflecting NLLB-200's English-centric training data. Despite this shift in absolute baseline values, the rank ordering of GRPO gains is highly consistent with English-source results (Spearman $\rho = 0.893$, $p = 0.007$, $n = 7$): Traditional Chinese remains the largest gainer in both conditions (+4.05 from Spanish vs. +3.94 from English), Arabic and Yoruba remain the smallest gainers (+0.12 and +0.11 from Spanish vs. +0.08 and +0.93 from English), and Tibetan retains a substantial gain (+2.71 vs. +3.01).

The Basque result deserves note: the Spanish gain (+0.94) is smaller than the English gain (+2.44) despite a lower Spanish baseline (44.63 vs. 47.75). Spanish→Basque is a more natural pair for NLLB-200 than English→Basque given the geographic and linguistic proximity, so the Spanish baseline is relatively stronger for Basque specifically, leaving less exploitable headroom. This is consistent with the headroom pattern rather than contradicting it.

		ΔchrF++		ΔC-22		ΔBS	
Group	**Lang.**	**600M**	**1.3B**	**600M**	**1.3B**	**600M**	**1.3B**
Aggl.	eus	+2.81	+2.57	+0.014	+0.011	+0.006	+0.006
	swh	+2.22	+3.00	+0.009	+0.012	+0.009	+0.011
	bod	+2.47†	+0.68†	+0.007	−0.001	+0.006	+0.002
	tur	+1.79	+1.69	+0.005	+0.002	+0.006	+0.005
	jpn	+0.48	+0.68	+0.015	+0.014	+0.011	+0.011
Isol.	zho_Hant	+3.45	**+4.88**	+0.086	+0.084	+0.109	**+0.106**
	zho_Hans	+1.41	+1.15	+0.030	+0.029	+0.037	+0.036
	yor	+0.36	−0.12	−0.001	−0.007	+0.002	+0.001
Root-pat.	arb	−0.13‡	+0.43	−0.002	+0.002	+0.001	+0.004
Fus.	bel	+0.37	+0.47	−0.004	+0.001	+0.001	+0.001
	ben	+1.68	+1.03	+0.004	+0.002	+0.006	+0.003
	ces	+0.96	+1.80	−0.001	+0.006	+0.009	+0.015
	pol	+0.42	+0.91	−0.002	+0.003	+0.001	+0.002

Table 8: Experiment B results (FLORES-200 devtest, *grpo_best* checkpoint). Training uses 10,000 monolingual English sentences from CCNews with no target-language supervision. All columns report deltas over the unfine-tuned NLLB baseline. ΔC-22 is the independent COMET-22 reference-based validator. ΔBS is the BERTScore F1 delta (`bert-base-multilingual-cased`, rescaled) against human references. **Bold** denotes the largest ΔchrF++ and ΔBS. † Tibetan gains are healthy at the best checkpoint but training without early stopping causes catastrophic collapse to below 2.0 chrF++ (§5). ‡ Arabic 600M shows marginal degradation at the best checkpoint; see §5.

Yoruba shows reward variance collapse at the final checkpoint (−3.43 chrF++) with best-checkpoint selection recovering a positive gain (+0.11), replicating the collapse mechanism observed for Tibetan in Experiment B (§5) and confirming that checkpoint selection is a necessary safeguard regardless of source language. Note that pooling Spanish and English observations to compute a single headroom correlation is not meaningful: absolute baseline values are source-language dependent (a Spanish baseline of 40 reflects different NLLB-200 behaviour than an English baseline of 40), so ranks across conditions are not directly comparable. The appropriate statistic is the rank correlation *between* conditions, reported above.

G Training Data Size Ablation

Table 3 reports the full data size ablation results for Basque (`eus_Latn`) and Traditional Chinese (`zho_Hant`) across $N \in \{100, 250, 500, 1{,}000\}$ training sentences. A reliable gain threshold emerges at $N = 500$: below this, neither SFT nor GRPO produces statistically significant improvements. At $N = 1{,}000$, GRPO significantly outperforms SFT for Basque ($p = 0.019$) but not for Traditional Chinese, where both methods approach the reward signal's discriminability ceiling.

H Catastrophic Forgetting Results

To assess whether LoRA fine-tuning on one language degrades NLLB-200's performance on unseen languages, each trained adapter is evaluated on three held-out languages: French (`fra_Latn`), Hindi (`hin_Deva`), and Russian (`rus_Cyrl`). A forgetting event is defined as a chrF++ degradation exceeding 1.0 point relative to the unfine-tuned baseline on the held-out language.

Across 39 GRPO adapter evaluations, zero forgetting events are observed. The distribution of chrF++ deltas on held-out languages is entirely positive or negligibly negative (range: −0.16 to +1.13, mean +0.34). The single forgetting event in the entire dataset occurs in the SFT condition: a Belarusian SFT adapter degrades Hindi by −0.75 chrF++. These results confirm that the LoRA + frozen base model configuration effectively preserves NLLB-200's other target language representations throughout GRPO training.

I Reward Discriminability Results

Table 11 reports Pearson r between reward score and quality rank for three representative LaBSE/COMET-Kiwi weight configurations across 12 languages, evaluated on a six-level quality cline of 50 sentences per language from the FLORES-200 development set. The cline consists of the human reference translation, four progressively degraded variants, and a nonsensical candi-

Group	Lang.	Base_{600M}	ΔchrF++ (600M)		ΔchrF++ (1.3B)	
			SFT_{3ep}	GRPO	$\text{Base}_{1.3B}$	GRPO
Aggl.	eus	45.32	+1.75	**+1.91**	46.40	**+2.07**
	swh	60.51	**+0.61**	−0.09	61.21	**+1.01**
	bod	24.60	+1.47	**+3.39**	26.81	**+0.98**
	tur	46.72	**+0.80**	+0.72	49.26	**+0.39**
	jpn	21.08	+1.03	**+2.08**	22.66	**+2.12**
Isol.	zho_Hant	7.60	+2.03	**+3.74**	7.75	**+5.22**
	zho_Hans	8.95	**+0.70**	+0.34	8.95	**+0.78**
	yor	16.55	**+2.13**	+0.78	18.11	**+0.79**
Root-pat.	arb	47.07	**+1.60**	+0.38	49.27	**+1.10**
Fus.	bel	50.36	−0.54	−0.63	53.20	−0.21
	ben	46.21	**+1.53**	+1.45	48.22	**+0.97**
	ces	52.10	**+0.10**	+0.00	54.77	**+0.20**
	pol	48.09	**+0.43**	+0.01	50.20	**+0.23**

Table 9: NTREX-128 evaluation (chrF++). All columns report deltas over the unfine-tuned NLLB baseline. GRPO is trained on CCNews monolingual source text (Experiment B); SFT_{3ep} is trained on FLORES-200 parallel data and is only available for the 600M model. **Bold** denotes the higher of SFT and GRPO within each model size. Belarusian (bel) degrades below baseline for both systems, indicating a domain mismatch independent of training objective.

Lang.	English (eng)		Spanish (spa)	
	Base	ΔchrF++	Base	ΔchrF++
zho_Hant	13.90	+3.94	10.43	**+4.05**
bod_Tibt	23.37	+3.01	22.24	+2.71
eus_Latn	47.75	+2.44	44.63	+0.94
ben_Beng	45.56	+1.98	37.17	+0.89
tur_Latn	52.68	+0.91	42.50	+0.82
yor_Latn	22.70	+0.93	21.50	+0.11
arb_Arab	51.12	+0.08	39.42	+0.12
$\rho = 0.893$, $p = 0.007$ (EN vs. ES rank)				

Table 10: GRPO ΔchrF++ on FLORES-200 devtest (600M) for English and Spanish source languages, sorted by English-source gain. Baselines differ between conditions because NLLB-200 performs differently on English→X vs. Spanish→X pairs. The rank ordering of gains is highly consistent across source languages ($\rho = 0.893$, $p = 0.007$), confirming that gain magnitude is driven by target-language properties rather than source-language training dynamics.

date. All configurations achieve good-to-excellent discrimination ($|r| > 0.81$ across all language-weight combinations), demonstrating robustness to the exact weighting choice.

The hybrid 0.50/0.50 configuration achieves the highest mean discrimination (−0.930) and is optimal for most morphologically rich languages. LaBSE-only is the best single-component configuration for Tibetan, Yoruba, Japanese, and both Chinese variants, consistent with weaker COMET-Kiwi calibration for languages underrepresented in its training data. For Arabic and Belarusian, the hybrid achieves the highest discrimination of any configuration, supporting the use of both components for root-pattern and Cyrillic fusional morphology.

Group	Lang.	LaBSE 1.00/0.00	Hybrid 0.50/0.50	CK 0.00/1.00
Aggl.	eus	−0.895	**−0.947**	**−0.947**
	tur	−0.888	−0.940	**−0.950**
	swh	−0.904	−0.945	**−0.948**
	bod	**−0.890**	−0.872	−0.119
	jpn	**−0.880**	−0.864	−0.836
Isol.	zho_Hant	**−0.892**	−0.885	−0.866
	zho_Hans	**−0.890**	−0.887	−0.877
	yor	**−0.942**	−0.929	−0.813
Root.	arb	−0.923	**−0.964**	−0.937
Fus.	bel	−0.906	**−0.958**	−0.940
	ben	−0.918	**−0.943**	−0.930
	ces	−0.897	**−0.955**	−0.923
	pol	−0.887	**−0.949**	−0.932
Mean (excl. bod)		−0.902	**−0.930**	−0.908

Table 11: Reward discriminability: Pearson r between reward score and quality rank on a six-level quality cline (50 sentences per language, FLORES-200 dev). **Bold** = best per language; ties both bolded. CK = COMET-Kiwi. Tibetan (bod) shows near-zero CK discriminability ($r = -0.119$); LaBSE-only strongly preferred. Mean excludes Tibetan for comparability with §5.

We leave a systematic sweep over reward weights during GRPO training to future work; the present analysis establishes that the equal-weight default is a principled starting point given the components' complementary coverage and comparable discriminative power.

J Decoding Regime Control Ablation

To assess whether GRPO gains reflect genuine policy improvement rather than decoding regime differences, we evaluate the unfine-tuned 600M baseline under temperature sampling ($T = 1.2$, matching GRPO training) on four languages. Temperature sampling of the untrained model substantially *degrades* chrF++ relative to beam search (Basque: $47.64 \rightarrow 30.03$; Arabic: $51.03 \rightarrow 27.86$). GRPO trained under the same sampling regime achieves $+2.93$ chrF++ over the beam search baseline for Basque and $+5.07$ for Traditional Chinese, substantially exceeding any decoding regime effect. The decoding regime therefore cannot account for the observed gains.

K Human Preference Evaluation

K.1 Setup

We evaluate four languages spanning a range of GRPO gain magnitudes: Traditional Chinese (zho_Hant, $+3.94$ chrF++), Bengali (ben_Beng, $+1.98$), Turkish (tur_Latn, $+0.91$), and Polish (pol_Latn, $+0.41$). These languages were selected to cover both the upper and lower ends of the gain distribution, allowing us to characterize not just whether gains are perceptible but at what magnitude they become so.

For each language, 50 source sentences were sampled uniformly from the FLORES-200 devtest set. Each sentence was paired with two translations: the Baseline NLLB-200 output (beam search, no fine-tuning) and the GRPO output. Annotators were presented with the source sentence and both translations in randomised order, with no indication of which system produced each output. They were asked to select the translation they considered higher quality, or to mark the pair as a tie if both were equivalent. No time limit was imposed. Annotation was conducted by bilingual speakers with native or near-native proficiency in the target language. Each sentence pair was evaluated by a single annotator.

K.2 Statistical Analysis

Preference counts were analysed using a one-sided binomial test against the chance hypothesis (equal preference for GRPO and Baseline among non-tied pairs). Ties were excluded from the binomial test following standard practice in MT preference evaluation (Graham et al., 2015). Sentences with missing annotations (annotator dropout) were excluded from all counts; effective sample sizes are reported in Table 12.

K.3 Results

Lang.	ΔchrF++	GRPO%	Base%	Tie%	p
zho_Hant	+3.94	68.0	8.0	24.0	<0.001
ben_Beng	+1.98	20.0	31.1	48.9	0.895
tur_Latn	+0.91	42.0	42.0	16.0	0.561
pol_Latn	+0.41	34.8	41.3	23.9	0.750

Table 12: Human preference results (50 sentences per language, 45–50 after excluding missing annotations). p: one-sided binomial test on non-tied pairs. ΔchrF++ from 600M, Experiment A.

Results reveal a clear perceptibility threshold. For Traditional Chinese, where GRPO achieves the largest automatic gain ($+3.94$ chrF++), annotators strongly and significantly prefer GRPO output (68% vs. 8% Baseline, $p < 0.001$). For the three remaining languages — Bengali, Turkish, and Polish, with gains of $+1.98$, $+0.91$, and $+0.41$ chrF++ respectively — preferences are indistinguishable from chance (all $p > 0.5$). This pattern is consistent with established findings that sub-2-point chrF++ improvements are typically below the threshold of human perceptibility (Graham et al., 2015).

Two aspects of the small-gain results deserve attention. First, Turkish and Polish show near-exact parity between GRPO and Baseline preference (42%/42% and 35%/41%) rather than a Baseline advantage, indicating the policy does not degrade perceived translation quality even when automatic metric gains are small. Second, the Bengali result — where Baseline is slightly preferred despite a $+1.98$ chrF++ gain — suggests that for moderate gains, GRPO may introduce changes that automatic metrics reward but that do not correspond to human-perceptible improvements; this is consistent with the known sensitivity of COMET-family metrics to surface fluency features that human judges do not weight equally.

K.4 Scope and Limitations

This evaluation is intentionally limited in scale. Each language was assessed by a single annotator on 50 sentences, with no inter-annotator agreement measurement. The goal was not to produce a definitive human evaluation but to perform a targeted sanity check: do the largest automatic gains

correspond to perceptible improvements? For Traditional Chinese the answer is unambiguously yes. For smaller gains, the evaluation is underpowered to draw strong conclusions, and a larger study with multiple annotators and inter-annotator agreement analysis would be needed to characterise the perceptibility threshold more precisely. We present these results as a frugal validation exercise rather than a substitute for full human evaluation, and note that the direction and pattern of results are consistent with what the headroom analysis would predict: gains large enough to matter are large enough to be noticed.