Calibrating Reward Design for English–Ukrainian Machine Translation

Anonymous ACL submission

Abstract

Reinforcement learning with verifiable rewards has driven recent progress in large language models, but its application to translation for morphologically rich, lower-resource languages remains underexplored. We present a systematic study of reward-driven fine-tuning with Group Relative Policy Optimization (GRPO) for English–Ukrainian translation. We propose *calibrated guardrail rewards*: structural signals gated by a normal-range function with parameters estimated from the FLoRes+ dev split. The guardrails saturate near 1.0 on well-formed outputs and decay only on outliers, freeing the gradient budget for core quality metrics. We pair this design with a multi-judge, multi-metric evaluation framework that combines three neural metrics with LLM judges from four independent model families. Across a controlled five-condition ablation on Lapa-12B, we show that calibrated guardrails outperform standard structural rewards on both directions, that the optimal reward composition differs sharply between en→uk and uk→en, and multi-judge approach highlights reward hacking of BLEU-based evaluation. An LLM-judge reward is the only configuration that improves both BLEU and judge score simultaneously on en→uk translation direction.

1 Introduction

Neural machine translation has progressed from supervised emulation of parallel corpora to alignment methods that fine-tune large language models with reinforcement learning (Ouyang et al., 2022; Shao et al., 2024; DeepSeek-AI et al., 2025). These methods have produced strong gains on high-resource language pairs, but their application to morphologically rich, lower-resource languages is comparatively unexplored. Ukrainian language, for example, has extensive nominal and verbal inflection, Cyrillic script, and is underrepresented in pre-training corpora, which makes it a hard target for systems trained primarily on web text. Recent progress in Ukrainian NMT has been driven by the Lapa LLM (Paniv et al., 2025), currently the highest-ranking system on the translation task of the Ukrainian LLM Leaderboard (Paniv, 2025), and by quality-scored parallel data at scale (Chaplynskyi and Zakharov, 2025). To our knowledge, no prior work has applied RL-based alignment specifically to Ukrainian translation.

A central obstacle in any such application is reward design. Unlike tasks with verifiable answers (Wen et al., 2025), translation is open-ended: many translations of the same source are valid, and any single metric is at best a proxy for quality. Nevertheless, combining multiple reward functions, introduces its own problem: signals that are active on every training step (binary format checks, length-ratio penalties, language-identification probes) compete with the core quality metrics for the gradient budget, even when the model's outputs are already well-formed.

This paper presents a systematic study of reward-driven fine-tuning with GRPO for English-Ukrainian translation. Our contributions are:

- **Calibrated guardrail rewards** gated by a normal-range function whose parameters are estimated from the FLoRes+ dev split. Each guardrail saturates near 1.0 on well-formed outputs and decays only on outliers, freeing the gradient budget for the core quality metrics. Adopting this design produces gains on both translation directions where previous structural reward mixes did not.
- **A controlled reward-component ablation** on Lapa-12B with nine conditions, isolating the effect of lexical anchors, neural Quality Estimation (QE) signals, standard versus calibrated guardrails, and an LLM-as-judge reward. The ablation reveals a sharp asymmetry between directions: en→uk requires cali-

brated guardrails or an LLM-judge reward to produce quality gains, while uk→en peaks at the simplest reward and degrades as components are added.

- **A multi-judge, multi-metric evaluation framework** that combines three neural metrics with LLM judges from four independent model families. The framework exposes cases where BLEU gains mask quality regressions and quantifies the self-bias that arises when judges come from the same family as the system under evaluation.

Code and trained adapters will be released upon acceptance.

2 Related Work

RL for translation. Early RL applications to NMT optimized PPO against neural quality metrics: Ramos et al. (2024) showed that PPO against COMET outperforms SFT and MBR decoding (Fernandes et al., 2022), and Ramos et al. (2025) extended this to token-level rewards via xCOMET error spans. Preference-based alternatives include DPO (Rafailov et al., 2024), CPO (Xu et al., 2024), and KTO (Ethayarajh et al., 2024), which avoid an explicit reward model but require curated preference data. More recently, GRPO has been applied to translation in MT-R1-Zero (Feng et al., 2025), DeepTrans (Wang et al., 2025), TAT-R1 (Li et al., 2025), and TranslateGemma (Google, 2025), using mixes of rule-based, neural-metric, and LLM-judge rewards. Alternative algorithms targeting reward composition (GDPO, Liu et al. 2026; MO-GRPO, Ichihara et al. 2026) and self-rewarding setups (Trans-Zero, Chai et al. 2025; SSR, Yang et al. 2025; RARO, Cai and Provilkov 2025) have been proposed but have not been evaluated on Ukrainian. We focus on a single algorithm (GRPO) and isolate the effect of reward design.

Verifiable rewards and reward hacking. For open-ended generation tasks, Wen et al. (2025) argue that verifiable signals are most robust. RLVRR (Jiang et al., 2026) extracts "reward chains" from references to verify content preservation; PB-RLSVR (Faisal et al., 2025) uses pivot-language back-translation as verification; NOVER (Liu et al., 2025) uses perplexity-derived self-rewards. WALAR (CMU, 2026) documents the specific failure modes of QE-based rewards in multilingual MT, where models learn degenerate outputs (source-copying, wrong-language output, over-translation) that score highly on CometKiwi-style metrics. Our calibrated guardrails and our observation that lexical anchors are required to prevent reward hacking under reference-free signals (Section 5, conditions F, G*, H) are consistent with WALAR's findings while operating at the reward-shape rather than rule-verifier level.

Ukrainian NLP and translation. Ukrainian-capable LLMs include Lapa LLM (Paniv et al., 2025), the highest-ranked system on the Ukrainian LLM Leaderboard (Paniv, 2025), and MamayLM (Yukhymenko et al., 2025). The FiftyFiveShades dataset (Chaplynskyi and Zakharov, 2025) provides 55M English–Ukrainian translation pairs scored by a QE ensemble that explains ∼60% of the variance in expert judgments. Dragoman (Paniv et al., 2024) previously demonstrated that SFT with k-fold perplexity filtering on ParaCrawl (Bañón et al., 2020) can lift a 7B model above NLLB-200 and GPT-4 on FLoRes. Despite this progress, RL-based alignment for Ukrainian remains, to our knowledge, unexplored.

3 Method

We fine-tune a pre-trained translation model with Group Relative Policy Optimization (GRPO; Shao et al. 2024) under a composite reward made up of three families: (i) lexical and reference-free quality metrics, (ii) calibrated guardrail signals that stay near saturation on well-formed outputs, and (iii) an optional LLM-as-judge reward. The first family is standard. The novelty is in the guardrail design: by gating each structural signal through a normal-range function whose parameters are estimated from a held-out development split, we keep these signals near 1.0 on in-distribution outputs and only let them shape the gradient when an output is anomalous.

3.1 GRPO Objective

For each source sentence x_i, the policy π_θ generates a group of G candidate translations $\{y_i^1, \ldots, y_i^G\}$. Each candidate receives a scalar reward

$$R(x_i, y_i^g) = \sum_{k=1}^{K} w_k \, r_k(x_i, y_i^g), \qquad (1)$$

the weighted sum of K component rewards. The advantage of each candidate is computed by nor-

malizing within the group:

$$\hat{A}_i^g = \frac{R(x_i, y_i^g) - \mu_i}{\sigma_i}, \tag{2}$$

where μ_i and σ_i are the group mean and standard deviation. The policy is updated with the clipped surrogate objective of Schulman et al. (2017), regularized by a KL penalty $\beta D_{\mathrm{KL}}[\pi_\theta \| \pi_{\mathrm{ref}}]$ that keeps π_θ close to the SFT reference. We use $G = 4$ throughout, which provides enough variance for meaningful advantage estimation while keeping generation cost manageable.

3.2 Quality Rewards

Three component rewards provide the core quality signal: chrF++ (Popovic, 2017), sentence-level BLEU (Papineni et al., 2002), and CometKiwi, the reference-free variant of COMET (Rei et al., 2020). The two lexical metrics measure surface fidelity to the reference and act as anchors that ground the policy in reference-faithful translations. CometKiwi is a reference-free QE metric trained on Direct Assessment data; for English–Ukrainian, it correlates above 0.60 with expert ratings (Chaplynskyi and Zakharov, 2025), making it usable as a learned reward. Lexical anchors and a neural QE signal constrain each other: the lexical pair prevents the model from drifting into fluent paraphrases that score well on CometKiwi but ignore the reference, while CometKiwi prevents BLEU-driven optimization from collapsing into strict word-for-word matches.

3.3 Calibrated Guardrail Rewards

Earlier work on metric-guided RL for translation (Ramos et al., 2024) includes structural rewards such as binary format checks, length-ratio penalties, and language-identification signals. In our preliminary runs, these active signals competed with the quality rewards for the gradient budget on every training step, even when the model's outputs were already well-formed. We replace them with *calibrated guardrails*, signals that are gated by a normal-range function so that they return a value near 1.0 when the output statistic is within the typical range observed on a held-out reference distribution, and decay exponentially when it deviates.

Each calibrated guardrail is defined by a parameter triple (μ, σ, n). We estimate μ and σ from the FLoRes+ dev split: for each (source, reference) pair we compute the raw statistic and take the sample mean and standard deviation of the resulting distribution. The third parameter n controls the width of the safe zone and is set to $n = 2$ throughout, so the safe zone covers roughly 95% of the reference distribution. For a given output statistic s, the guardrail reward is

$$r_{\mathrm{cal}}(s) = \min\Big(1,\ \exp\Big(n - \tfrac{|s-\mu|}{\sigma}\Big)\Big). \tag{3}$$

Inside the safe zone, r_{cal} is exactly 1.0 and contributes no gradient. Outside, it decays exponentially in proportion to how many standard deviations s lies from μ. Because well-formed candidates sit at the saturation value, the gradient budget for these components is spent only on outliers (all-caps generation, missing punctuation, runaway repetition, untranslated source segments).

We define five guardrails. The first four use Equation 3 on the signed difference between source-side and hypothesis-side ratios; the fifth (Equation 4) is a multiset overlap score that already lies in $[0, 1]$ and does not need calibration:

1. **Uppercase ratio**: difference in the ratio of uppercase characters to total characters.
2. **Special characters ratio**: difference in the ratio of non-letter, non-space characters.
3. **Word count ratio**: difference in words per character, detecting truncation and run-on generation.
4. **Unique words ratio**: difference in unique-words per character, detecting degenerate repetition.
5. **Unpaired special characters**:

$$r_{\mathrm{overlap}} = 1 - u/t, \tag{4}$$

where u is the multiset symmetric difference of special characters between source and translation and t is their combined count.

Each guardrail receives weight 0.05, totaling 0.25, and the remaining 0.75 is allocated to quality rewards. The calibration parameters differ between translation directions: μ flips sign across directions because Ukrainian text uses more words per character than English (Appendix A). Using parameters from the wrong direction makes the word-count guardrail overly active ($\sim$0.65 instead of $\sim$0.99) and wastes the gradient budget.

3.4 LLM-as-Judge Reward

The quality rewards in Section 3.2 capture surface fidelity and reference-free fluency but are blind to translation errors that require deeper semantic understanding, such as mistranslation and incorrect terminology. We add an LLM-as-judge reward: for each (source, reference, hypothesis) triple, a multilingual evaluator outputs an integer Direct Assessment score on a 0–100 scale following the GEMBA-DA protocol (Kocmi and Federmann, 2023b), which is normalized to $[0, 1]$. We use M-Prometheus-14B (Kim et al., 2024), a Qwen2.5-14B model fine-tuned for multilingual evaluation. When the judge reward is included, it carries weight 0.60, with chrF++ and BLEU reduced to 0.10 and 0.05 respectively and the five guardrails keeping their 0.05 weights each.

4 Experimental Setup

Model. We use Lapa v0.1.2 Instruct (Paniv et al., 2025), a 12.2B-parameter model based on Gemma-3 (Team et al., 2025) that has been further pre-trained on Ukrainian data and instruction-tuned. Lapa handles both translation directions with a single set of weights. We fine-tune it with LoRA (Hu et al., 2022) of rank $r = 128$ and $\alpha = 256$ on the attention projections (q/k/v/o). An earlier $r = 32$ configuration showed no improvement over the baseline.

Training data. The training corpus is WikiMatrix (Schwenk et al., 2019), 132K English–Ukrainian sentence pairs mined from Wikipedia.

Benchmarks. We evaluate on FLoRes+ devtest (Goyal et al., 2022), 1,012 Wikipedia-derived sentence pairs, and on WMT24++ (Kocmi et al., 2024), 960 pairs spanning news, social media, speech transcripts, and literary text. The FLoRes+ domain overlaps with WikiMatrix; WMT24++ is used to test generalization. All evaluation uses the prompt *"You are a professional translator. You give only the translated text and nothing else. Translate the following text into {lang}:\n{text}"*, greedy decoding, batched inference with batch size 16. Configurations are held constant across all comparisons.

Automatic metrics. We report BLEU (Papineni et al., 2002) and chrF++ (Popovic, 2017) for surface fidelity; CometKiwi-DA-XL, xCOMET-XXL (Guerreiro et al., 2023), and MetricX-23-XL (Juraska et al., 2024) for learned neural quality. MetricX is inverted in all figures so that higher is better across metrics.

LLM-as-judge evaluation. To avoid single-judge bias, we evaluate each system with judges from four independent model families under the GEMBA-DA protocol (Kocmi and Federmann, 2023b): Mistral-Small-3.2-24B (primary, fully independent from all evaluated models), M-Prometheus-14B (Kim et al., 2024), Aya-Expanse-32B (for AI, 2024), and a Gemma-4-31B GEMBA-MQM annotator (Kocmi and Federmann, 2023a) for error-category breakdowns. The first three produce 0–100 Direct Assessment scores; the last produces span-level error annotations with severity labels.

Statistical testing. We validate all claims with two paired tests on per-sentence judge scores: a Wilcoxon signed-rank test (two-sided p-values, threshold $p < 0.05$) and a 1,000-resample bootstrap 95% confidence interval on the mean delta. A delta is treated as significant when its 95% CI excludes zero. Because both LLM judges output integer scores on a coarse grid, 82–95% of source sentences receive identical scores under baseline and trained model, so the Wilcoxon test has limited power on per-cell comparisons; we rely on agreement across judges and neural metrics as the primary evidence of improvement (Appendix C).

Training infrastructure. All experiments run on 4× NVIDIA RTX 6000 Ada (48 GB). For Lapa, we use a split-GPU setup with vLLM (Kwon et al., 2023) on two GPUs and DeepSpeed ZeRO-2 (Rajbhandari et al., 2020) with CPU optimizer offload on the others, achieving ∼14 s per step. Training uses TRL (von Werra et al., 2022) with learning rate 5×10^{-6}, $\beta = 0.1$, batch 2×8, 300 steps, and group size $G = 4$, held constant across all conditions.

5 Reward Ablation

We isolate the contribution of each reward component with nine conditions trained on Lapa with identical hyperparameters (Section 4) and differing only in reward composition (Table 1). Conditions A–E vary the lexical anchors, the neural QE signal, the choice between standard and calibrated guardrails, and the addition of the LLM-judge reward. Conditions F, G, G*, and H replace the lexical anchors entirely with a neural QE metric as the sole quality signal while keeping calibrated guardrails fixed;

Table 1: Ablation conditions. Numbers are reward weights summing to 1.0. "cal" is the five calibrated guardrails at 0.05 each (0.25 total); "std" is the standard format/length/language triple at 0.05 each (0.15 total). G uses xCOMET in reference-based unified mode; G* in QE mode plus a 0.10 language-ID weight.

Cond.	chrF	BLEU	Neural QE	Guards	Judge
A	.55	.45	–	–	–
B	.30	.25	Kiwi .45	–	–
C	.30	.25	Kiwi .20	std	–
D	.30	.25	Kiwi .20	cal	–
E	.10	.05	–	cal	.60
F	–	–	Kiwi .75	cal	–
G	–	–	xCOMET-r .75	cal	–
G*	–	–	xCOMET-QE .65	cal	–
H	–	–	MetricX .75	cal	–

they test whether reference-free training can substitute for chrF and BLEU.

Each condition is evaluated on FLoRes+ and WMT24 in both directions with BLEU and the Mistral-Small judge. Table 2 reports Δ values against the Lapa baseline, and Figure 1 plots BLEU against the Mistral judge per condition, with the lower-right quadrant marking the reward-hacking regime.

en→uk is prone to reward hacking. Condition A increases FLoRes BLEU by $+0.30$ but decreases the Mistral judge by a significant -0.25. Condition C, with standard structural guardrails, shows a worse pattern: -0.56 BLEU and a significant -0.21 judge regression. Only the calibrated guardrails (D) and the LLM-judge reward (E) yield positive judge deltas on en→uk on FLoRes+. Condition E is the only configuration where BLEU ($+0.58$) and judge ($+0.12$) improve simultaneously on FLoRes en→uk: the LLM judge accesses quality dimensions (mistranslation, terminology) that lexical metrics miss.

uk→en benefits from simple rewards. The simplest condition (A) produces the largest uk→en gains on both benchmarks: $+1.13$ FLoRes BLEU with a significant $+0.23$ judge improvement, and $+1.15$ WMT24 BLEU. Every additional reward component dilutes this effect.

Calibrated vs. standard guardrails. Comparing C (standard guardrails) and D (calibrated guardrails) with the same underlying quality mix isolates the effect of the calibration design. On en→uk, the difference is $+1.07$ BLEU and $+0.30$ judge on FLoRes, and -0.54 BLEU but $+0.28$ judge on WMT24. D is also the only condition that achieves a statistically significant judge improvement on WMT24 en→uk ($+0.37$). The same calibrated guardrails also bring B's judge regression to a positive on uk→en's WMT24 cell.

Neural QE alone leads to reward hacking. Conditions F (CometKiwi), G* (xCOMET-QE), and H (MetricX) replace the lexical anchors with a neural QE metric. All three degrade BLEU on every benchmark–direction cell and degrade the Mistral judge in seven of twelve. The pattern is most visible on the metric used as the reward itself: H improves MetricX by -0.22 on WMT24 en→uk (lower is better) while losing BLEU and judge, the textbook reward-hacking signature. Condition G is the exception: keeping the reference inside xCOMET's unified scorer anchors the output and avoids drift, but produces no gain. This confirms across all three neural-metric families that lexical anchors are essential; reference-free neural metrics alone are reward-hackable regardless of the metric.

SFT degrades the strong baseline. Supervised fine-tuning on WikiMatrix (one epoch, cross-entropy) degrades judge scores on all four benchmark–direction cells. On uk→en the SFT row exhibits its own form of reward hacking: BLEU increases ($+0.83$ FLoRes, $+1.18$ WMT24) while judge scores fall (-0.90, -0.76). Cross-entropy training forces a model already above the reference's quality ceiling toward those references, which is visible as a distribution shift.

6 Multi-Judge Analysis

The ablation reveals reward hacking on en→uk where BLEU improvements mask quality regressions. We cross-validate this with five independent quality metrics (CometKiwi, xCOMET, MetricX, Aya, Mistral) and document two further phenomena visible only with multi-judge evaluation: judge family bias and an absolute-score anomaly in CometKiwi.

Multi-metric confirmation of reward hacking. On FLoRes en→uk, condition A improves BLEU by $+0.30$. Every other quality metric disagrees: CometKiwi -0.0056, xCOMET -0.0069, MetricX $+0.08$ (lower is better), Aya judge -0.07, Mistral judge -0.25. Five of five quality signals from three metric families agree that condition A degrades en→uk translation quality. The same divergence pattern appears in condition C on FLoRes

Table 2: Reward ablation on Lapa: deltas vs. the Lapa baseline (FLoRes en→uk BLEU 33.45; FLoRes uk→en BLEU 42.00; WMT24 en→uk BLEU 31.21; WMT24 uk→en BLEU 34.54). Judge is the Mistral-Small 0–100 score. Significant Wilcoxon $p < 0.05$ marked with * (per-cell table in Appendix C). Best Δ per direction–benchmark column in **bold**.

	FLoRes+				WMT24			
	en→uk		uk→en		en→uk		uk→en	
Cond.	ΔBLEU	ΔJ	ΔBLEU	ΔJ	ΔBLEU	ΔJ	ΔBLEU	ΔJ
A	+0.30	−0.25*	**+1.13**	**+0.23***	−0.19	−0.12	**+1.15**	**+0.15**
B	+0.04	−0.06	+0.84	+0.14*	−0.02	+0.09	+0.86	+0.09
C	−0.56	−0.21*	+0.74	+0.05	**+0.93**	+0.09	+0.85	+0.09
D	+0.51	+0.09	+0.48	+0.04	+0.39	**+0.37***	+0.35	+0.13
E	**+0.58**	**+0.12**	+0.26	−0.07	+0.04	+0.22	+0.14	−0.01
F	−0.46	−0.16	−0.69	−0.07	−1.03	+0.02	−0.77	−0.43
G	+0.43	+0.04	−0.86	−0.04	−0.48	+0.05	−0.80	−0.04
G*	+0.42	−0.10	−1.58	−0.36	−0.51	−0.07	−0.95	−0.15
H	−0.67	−0.16	−0.92	+0.04	−1.19	−0.05	−0.93	−0.14
SFT	−2.15	−1.66*	+0.83	−0.90*	−2.69	−2.10*	+1.18	−0.76*

en→uk, and on the neural-metric-only conditions F and H, where the metric used as the reward improves while BLEU and judges fall.

Judge self-bias. Comparing Lapa+GRPO (condition E) against Qwen3.5-122B on the same prompts isolates judge family bias. BLEU favors Lapa by -3.5 to -6.3 points on every benchmark–direction cell. Qwen-122B judging itself reverses the verdict by $+0.77$ to $+1.86$ points across benchmarks. M-Prometheus-14B, a Qwen2.5-14B fine-tune, retains residual family bias of $+0.37$ to $+1.12$. Mistral-Small, the only judge from a fully independent family, exhibits the smallest gap and is the only judge that agrees with BLEU on FLoRes en→uk. This is why we use Mistral as the primary judge throughout the paper, and why no single-judge claim, regardless of which judge, should be reported in isolation.

Cross-metric agreement varies by benchmark. We compute system-level Spearman rank correlations between metrics over the 20 systems for which all metrics are available (Lapa baseline, the nine ablation conditions, the SFT checkpoint, and a panel of open-weights models from Appendix B, with MetricX inverted so that higher means better quality). Three metric families emerge: lexical (BLEU, chrF), neural (CometKiwi, xCOMET, MetricX), and LLM judges (Aya, Mistral). Agreement is strongest on WMT24 en→uk ($\rho \geq 0.85$ within the COMET family and between Aya and Mistral) and weakest on FLoRes uk→en (cross-family ρ as low as -0.64). With $n = 20$, individual correlations below $|\rho| \approx 0.45$ are not significant at $p < 0.05$, so we treat the family structure and the WMT24/FLoRes contrast, not individual cells, as the supporting observations.

CometKiwi exceeds the human reference. On FLoRes en→uk, CometKiwi scores the Lapa baseline at 0.7915, higher than the 0.7746 it assigns to the human reference. CometKiwi cannot therefore be used to estimate remaining headroom: it would suggest the model already exceeds the human ceiling, which the LLM judges and the MQM error analysis (Appendix D) contradict. Combined with the F/H rows in Table 2, this argues against using CometKiwi as a sole reward signal or as a ceiling estimator without external validation.

7 Conclusion

We presented a controlled study of reward-driven GRPO fine-tuning for English–Ukrainian translation. Calibrated guardrail rewards, structural signals gated by a normal-range function calibrated on a held-out dev split, turn standard structural penalties into passive safety nets that free the gradient budget for the core quality metrics. On Lapa-12B, calibrated guardrails (condition D) are the first configuration to produce a positive Mistral-judge delta on en→uk, and the only configuration with a statistically significant judge improvement on WMT24 en→uk. Pairing calibrated guardrails with an LLM-as-judge reward (condition E) is the only configuration where BLEU and judge improve simultaneously on FLoRes en→uk. The multi-judge, multi-metric framework around the ablation makes two findings visible that single-metric reporting cannot: that "best BLEU" is not always best translation (re-

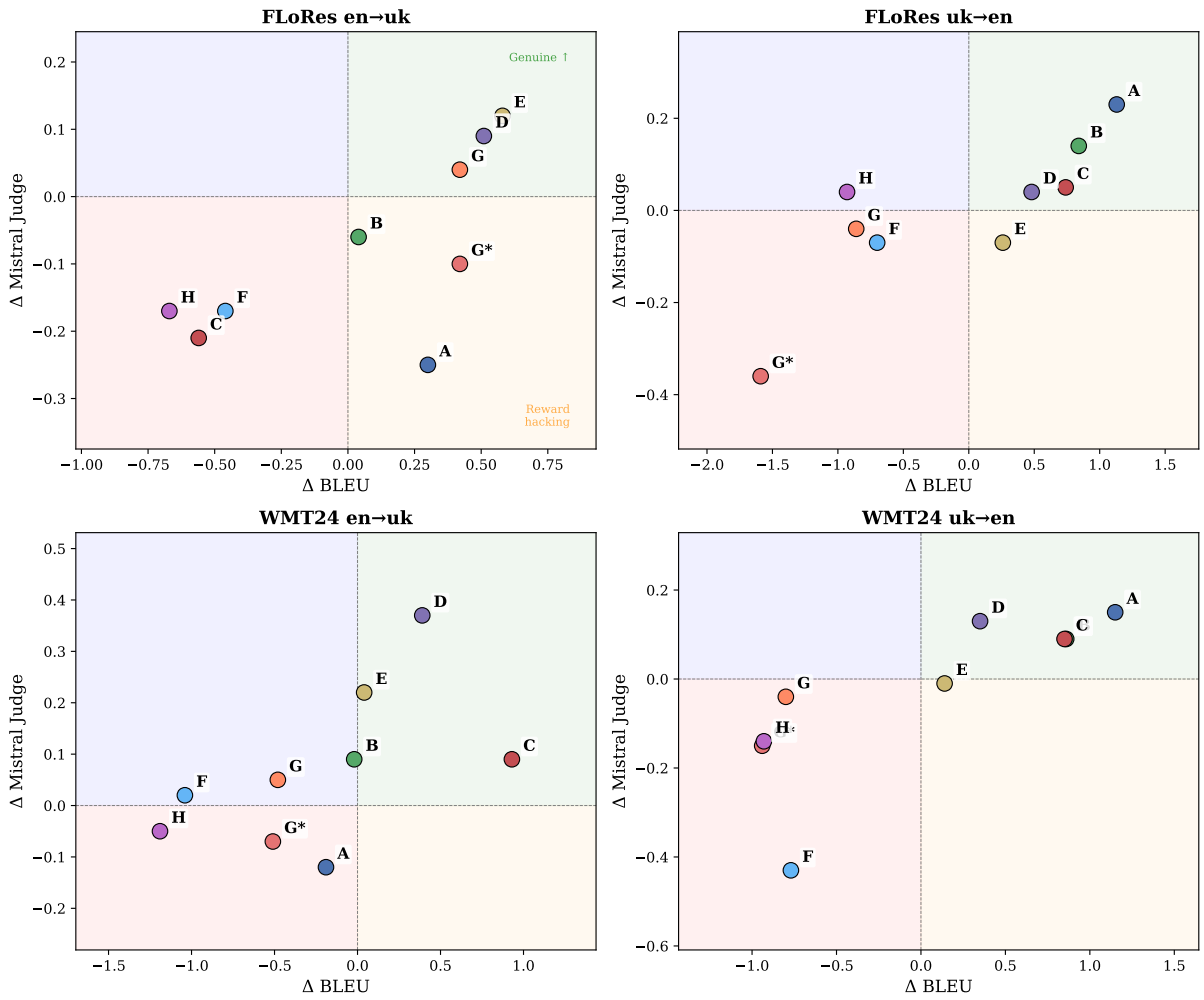

Figure 1: Per-condition ΔBLEU vs. ΔMistral-judge against the Lapa baseline, one panel per benchmark–direction. The lower-right quadrant marks reward hacking (BLEU improves while judge regresses); the upper-right marks joint improvement on both axes. On FLoRes en→uk the simplest conditions (A, C) sit in the reward-hacking quadrant; only D and E land in the upper-right quadrant. On uk→en the picture inverts: A produces the largest gains and added components drift the points downward.

ward hacking on en→uk; SFT on uk→en), and that the optimal reward composition is asymmetric between directions.

Limitations

Domain specificity. The improvements reported here are tied to the training corpus. WikiMatrix and FLoRes+ share a Wikipedia domain, and en→uk gains on the out-of-domain WMT24 benchmark are consistently smaller than on FLoRes+. The more general claim that calibrated guardrails work in any domain is not supported by our data.

No human evaluation. We replaced single-metric evaluation with a multi-judge, multi-metric framework, but did not conduct human evaluation. The multi-judge design partially compensates: four judges from independent families agreeing on a finding is stronger evidence than any single automatic metric. However, cross-metric correlations vary substantially by benchmark (system-level Spearman ρ as low as -0.64 on FLoRes uk→en), and some condition–cell results are metric-dependent.

Limited Wilcoxon power on per-cell tests. Because the LLM judges output integer scores on a 0–100 scale, 82–95% of source sentences receive identical scores under baseline and trained model, leaving only a small fraction of pairs with non-zero deltas. The per-cell Wilcoxon signed-rank test has limited power to detect modest mean shifts, and non-significant positive cells in Table 2 should be read as "insufficient power on this judge" rather than "no effect." We mitigate this by relying on

agreement across multiple judges and neural metrics rather than per-cell significance as the primary evidence of improvement.

Residual reward hacking. Even with calibrated guardrails, reward hacking remains a risk on en→uk. While the LLM-judge reward (condition E) mitigates it on FLoRes+, the benefit does not generalize to WMT24.

Data scale and GRPO headroom. WikiMatrix is relatively small (132K pairs) for RL fine-tuning, and we observe a consistent pattern in which GRPO improvement shrinks as baseline quality increases. The method's applicability is bounded by the signal available in the training data and by the reward functions; whether the same effects hold at 30B+ scale or on substantially larger training corpora is not established by our experiments.

Single model family and language pair. We study one base model (Lapa-12B) and one language pair (English–Ukrainian). The asymmetry we observe between translation directions, the calibration parameters, and the relative contributions of each reward component may differ for other low-resource pairs or other base models.

Ethical Considerations

Use of LLM judges. Using LLM judges as a reward signal during training and as evaluators of the same systems risks circular reasoning. We mitigate this in two ways: the training-time judge (M-Prometheus-14B) is from a different family than the primary evaluation judge (Mistral-Small-3.2-24B), and our multi-judge framework quantifies family bias (Section 6). Single-judge evaluation, regardless of which judge, can produce systematically biased rankings between models from the same family.

Compute and environmental cost. All experiments ran on 4× NVIDIA RTX 6000 Ada (49 GB). A full 300-step Lapa training takes ∼1.2 hours; the full ablation in this paper, including baselines, repeated reward configurations, and judge evaluation passes, accounted for roughly 1,200 GPU-hours.

AI-assisted authoring. AI coding assistants (Google's Antigravity IDE; Anthropic's Claude Code) were used for code-writing, debugging, and organizing experiment logs across runs, and for style and grammar correction on the text of this paper. All experimental designs, reward formulations, analyses, and the ideas and claims in this paper are the authors' own. Tools were used as productivity aids; all code and artifacts were reviewed by the authors before being reported.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Locke Cai and Ivan Provilkov. 2025. Escaping the verifier: Learning to reason via demonstrations. *Preprint*, arXiv:2511.21667.

Yekun Chai and 1 others. 2025. Trans-zero: Self-play translation via llm. *Preprint*, arXiv:2504.14669.

Dmytro Chaplynskyi and Kyrylo Zakharov. 2025. A framework for large-scale parallel corpus evaluation: Ensemble quality estimation models versus human assessment. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 73–85, Vienna, Austria (online). Association for Computational Linguistics.

CMU. 2026. Mending the holes: Word alignment for rl in machine translation. *Preprint*, arXiv:2603.13045.

DeepSeek-AI and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

Fahim Faisal, Kaiqiang Song, Song Wang, Simin Ma, Shujian Liu, Haoyun Deng, and Sathish Reddy Indurthi. 2025. Aligning multilingual reasoning with verifiable semantics from a high-resource expert model. *Preprint*, arXiv:2509.25543.

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, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and André F. T. Martins. 2022. Quality-aware decoding for neural machine translation. *Preprint*, arXiv:2205.00978.

Cohere for AI. 2024. Aya expanse: Connecting the multilingual world through ai.

Google. 2025. Translategemma: Translation-specialized gemma models. *Preprint*, arXiv:2601.09012.

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-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhi Chen. 2022. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Mitsuki Sakamoto, Ryota Mitsuhashi, and Eiji Uchibe. 2026. Mo-grpo: Mitigating reward hacking of group relative policy optimization on multi-objective problems. *Preprint*, arXiv:2509.22047.

Yuxin Jiang, Yufei Wang, Qiyuan Zhang, Xingshan Zeng, Liangyou Li, Jierun Chen, Chaofan Tao, Haoli Bai, and Lifeng Shang. 2026. From verifiable dot to reward chain: Harnessing verifiable reference-based rewards for reinforcement learning of open-ended generation. In *The Fourteenth International Conference on Learning Representations*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, and 2 others. 2024. Preliminary wmt24 ranking of general mt systems and llms. *Preprint*, arXiv:2407.19884.

Tom Kocmi and Christian Federmann. 2023a. Gemba-mqm: Detecting translation quality error spans with gpt-4. *Preprint*, arXiv:2310.13988.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. *Preprint*, arXiv:2302.14520.

Woosuk Kwon, Zhuohan Li, Sicheng Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Zheng Li, Mao Zheng, Mingyang Song, and Wenjie Yang. 2025. Tat-r1: Terminology-aware translation with reinforcement learning and word alignment. *Preprint*, arXiv:2505.21172.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. 2026. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *Preprint*, arXiv:2601.05242.

Wei Liu, Siya Qi, Xinyu Wang, Chen Qian, Yali Du, and Yulan He. 2025. Nover: Incentive training for language models via verifier-free reinforcement learning. *Preprint*, arXiv:2505.16022.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Yurii Paniv. 2025. Isolating llm performance gains in pre-training versus instruction-tuning for mid-resource languages: The ukrainian benchmark study. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 876–883, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yurii Paniv, Dmytro Chaplynskyi, Nikita Trynus, and Volodymyr Kyrylov. 2024. Setting up the data printer with improved english to ukrainian machine translation. *Preprint*, arXiv:2404.15196.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. Lapa LLM v0.1.2: the most efficient Ukrainian open-source language model.

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.

Maja Popovic. 2017. chrf ++ : words helping character n-grams.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Samyam Rajbhandari, Jeff Rasley, Olatunji Rinks, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Miguel Moura Ramos, Tomás Almeida, Daniel Vareta, Filipe Azevedo, Sweta Agrawal, Patrick Fernandes, and André F. T. Martins. 2025. Fine-grained reward optimization for machine translation using error severity mappings. *Preprint*, arXiv:2411.05986.

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and André F. T. Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. *Preprint*, arXiv:2311.09132.

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

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

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *Preprint*, arXiv:1907.05791.

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.

Gemma Team and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2022. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025. Deeptrans: Deep reasoning translation via reinforcement learning. *Preprint*, arXiv:2504.10187.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *Preprint*, arXiv:2506.14245.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Preprint*, arXiv:2401.08417.

Yang and 1 others. 2025. Ssr-zero: Simple self-rewarding rl for mt. *Preprint*, arXiv:2505.16637.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm: An efficient state-of-the-art ukrainian llm. https://huggingface.co/blog/INSAIT-Institute/mamaylm.

A Calibration Parameter Estimation

The calibrated guardrails in Appendix 3.3 are parameterized by a tuple (μ, σ, n) per signal. We estimate μ and σ from the sentence-aligned (source, reference) pairs of the FLoRes+ dev split (997 pairs per direction). For each pair, we compute the raw statistic (the signed difference between source-side and hypothesis-side ratios for that signal), and take the sample mean and standard deviation. The third parameter n is set to 2 throughout, so the safe zone covers $\sim$95% of the reference distribution and only outliers receive a non-saturating reward. Table 3 reports the estimated values. The mean μ flips sign between directions because Ukrainian text uses more words per character than English: the word-count ratio has $\mu = +0.025$ for en→uk but $\mu = -0.025$ for uk→en. Using parameters estimated for the wrong direction makes the word-count guardrail overly active during training ($\sim$0.65 instead of $\sim$0.99 on in-distribution outputs), consuming gradient signal that should flow to the core quality rewards.

Table 3: Calibration parameters for the four calibrated guardrails, estimated from FLoRes+ dev. All guardrails use $n = 2$.

	en→uk		uk→en	
Signal	μ	σ	μ	σ
Uppercase ratio	$+.005$	.011	$-.005$	.011
Special chars ratio	$-.005$	.014	$+.005$	.014
Word count ratio	$+.025$	.018	$-.025$	.018
Unique words ratio	$+.014$	.018	$-.014$	.018

The fifth guardrail (unpaired special characters) is a multiset overlap score (Equation 4) that already lies in $[0, 1]$ and does not require calibration.

B Open-Weights Model Comparison

To contextualize the Lapa+GRPO results, we benchmark a panel of recent open-weights LLMs on

the same four benchmark–direction cells, using the identical professional-translator prompt and greedy decoding. Lapa+GRPO (best) uses the best-performing GRPO condition per direction: E on FLoRes en→uk, A on FLoRes uk→en, D on WMT24 en→uk, and A on WMT24 uk→en. Lapa+GRPO (merged) is the deployable single-model linear combination of the E (en→uk) and A (uk→en) adapters. Bold marks the best value per column; MetricX is lower-is-better. Tables 4–7 report BLEU, chrF, CometKiwi (CK), xCOMET, MetricX, and the Mistral-Small and Aya-Expanse judges.

Table 4: FLoRes+ en→uk.

Model	BLEU	chrF	CK	xCOMET	MetricX↓	Mistral	Aya
Lapa+GRPO best (E)	**34.02**	**61.63**	0.789	0.936	0.83	83.94	92.00
Lapa+GRPO (merged)	33.88	61.54	0.789	0.934	0.84	83.87	91.94
Lapa baseline	33.50	61.24	0.792	0.939	0.79	83.82	91.85
Gemma-4-31B	30.96	60.20	0.797	**0.949**	**0.63**	**84.59**	**92.36**
Gemma-4-26B-A4B	30.40	60.04	**0.800**	0.947	0.68	84.31	92.30
Gemma-4-e4B	30.86	59.24	0.780	0.913	0.98	82.61	90.97
Gemma-3-12B-it	28.36	57.79	0.789	0.931	0.81	82.54	91.27
Qwen-3.6-27B	29.79	58.98	0.795	0.937	0.76	83.61	91.82
Qwen-3.6-35B-A3B	28.11	58.13	0.798	0.940	0.72	83.41	91.67
GLM-4.7-Flash	25.29	55.36	0.767	0.903	1.00	80.64	89.60

Table 5: FLoRes+ uk→en.

Model	BLEU	chrF	CK	xCOMET	MetricX↓	Mistral	Aya
Lapa+GRPO best (A)	43.15	67.85	**0.766**	0.947	1.35	88.62	94.33
Lapa+GRPO (merged)	**43.28**	**67.97**	0.766	0.947	1.35	88.59	94.32
Lapa baseline	42.02	67.39	0.766	0.948	1.37	88.46	94.25
Gemma-4-31B	42.45	67.74	0.766	**0.951**	1.32	**89.12**	**94.53**
Gemma-4-26B-A4B	40.89	67.01	0.764	0.947	1.35	88.72	94.31
Gemma-4-e4B	39.94	65.68	0.760	0.931	1.51	87.16	93.58
Gemma-3-12B-it	40.30	66.25	0.762	0.938	1.36	88.45	94.28
Qwen-3.6-27B	40.62	66.68	0.765	0.948	1.32	88.92	94.37
Qwen-3.6-35B-A3B	39.41	65.82	0.764	0.948	**1.30**	88.69	94.38
GLM-4.7-Flash	39.59	65.42	0.760	0.934	1.45	87.68	93.76

Table 6: WMT24 en→uk.

Model	BLEU	chrF	CK	xCOMET	MetricX↓	Mistral	Aya
Lapa+GRPO best (D)	**31.59**	**57.35**	0.694	0.801	2.46	80.59	89.10
Lapa+GRPO (merged)	31.37	57.15	0.691	0.795	2.53	80.31	88.88
Lapa baseline	31.12	56.97	0.700	0.806	2.36	80.22	88.76
Gemma-4-31B	30.44	57.27	**0.709**	**0.833**	**2.08**	**81.50**	**89.53**
Gemma-4-26B-A4B	30.35	57.31	0.706	0.831	2.10	81.00	89.46
Gemma-4-e4B	28.99	55.22	0.678	0.779	2.68	79.73	88.42
Gemma-3-12B-it	27.89	54.51	0.685	0.799	2.36	79.35	88.17
Qwen-3.6-27B	28.00	55.43	0.704	0.824	2.12	80.40	88.98
Qwen-3.6-35B-A3B	27.84	55.07	0.700	0.815	2.12	80.56	88.96
GLM-4.7-Flash	23.93	51.32	0.663	0.758	2.70	77.23	86.56

C Wilcoxon Power Analysis

The Wilcoxon signed-rank test in Appendix 4 compares per-sentence judge scores between baseline and trained systems. Because both LLM judges output integer scores on a coarse 0–100 grid (Mistral-Small effectively uses 5-point increments, Aya-Expanse 10-point), the large majority of source sentences receive identical scores under the baseline and the trained model, contributing zero to the rank

Table 7: WMT24 uk→en.

Model	BLEU	chrF	CK	xCOMET	MetricX↓	Mistral	Aya
Lapa+GRPO best (A)	**35.75**	60.70	0.705	0.821	2.59	84.74	91.84
Lapa+GRPO (merged)	35.69	60.71	0.705	0.820	2.60	84.62	91.75
Lapa baseline	34.60	60.17	0.707	0.825	2.59	84.59	91.87
Gemma-4-31B	35.18	**61.02**	**0.709**	**0.838**	**2.48**	**85.71**	**92.54**
Gemma-4-26B-A4B	33.94	60.35	0.706	0.832	2.52	85.18	92.42
Gemma-4-e4B	33.42	59.12	0.703	0.807	2.76	83.70	91.21
Gemma-3-12B-it	33.45	59.19	0.699	0.815	2.59	84.93	91.99
Qwen-3.6-27B	31.96	58.91	0.708	0.833	2.48	85.33	92.01
Qwen-3.6-35B-A3B	31.76	58.72	0.707	0.829	2.51	85.26	92.09
GLM-4.7-Flash	31.97	58.43	0.697	0.803	2.73	83.48	91.09

statistic. Empirically, 82–95% of pairs are scored identically across the ablation conditions and benchmarks, leaving only 5–18% of pairs that actually carry signal.

With so few non-zero deltas, the test has limited power to detect modest shifts in mean judge score, even when the cross-metric and cross-judge picture clearly points to a real effect. For example, condition D produces a positive judge delta on FLoRes en→uk (+0.09 on Mistral), in agreement with two other neural metrics, but the per-cell Wilcoxon test does not reach $p < 0.05$ on this single judge. Treating such cells as "no effect" would be the wrong inference; we read them as "insufficient power on this judge."

The cells that do reach significance in Table 2 of the main paper are: condition A on FLoRes en→uk (judge −0.25) and FLoRes uk→en (judge +0.23); condition B on FLoRes uk→en (judge +0.14); condition C on FLoRes en→uk (judge −0.21); condition D on WMT24 en→uk (judge +0.37); and SFT on all four cells (judge −0.76 to −2.10). SFT's large negative effect size makes it the easiest to detect; the GRPO cells that reach significance are precisely those with the largest absolute deltas. This is the expected behavior of a paired test on coarse ordinal data, and is why our claims rely on agreement across judges and neural metrics rather than per-cell significance.

D Qualitative Examples

Examples below are not hand-picked. For each benchmark–direction–condition triple, we rank devtest sentences by per-sentence Mistral-Small judge delta and report the largest improvement, the largest regression, and a near-zero case. Pairs where baseline and trained outputs are character-identical are skipped, since any judge delta on those is noise. Baseline-only spans are shown in red; trained-only spans are shown in green.

FLoRes+ en→uk, baseline vs. condition E. *Largest improvement* (judge $\Delta = +20$, $65 \rightarrow 85$):

Src The vertical clearance under the bridge is 15 meters.
Ref Вертикальний просвіт під мостом становить 15 метрів.
Base Висота над мостом становить 15 метрів.
GRPO Вертикальний просвіт під мостом становить 15 метрів.

The trained model produces the precise technical term *vertical clearance*, where the baseline uses the generic *height above*.

Largest regression (judge $\Delta = -25$, $75 \rightarrow 50$):

Src The show originally featured amateur voice actors.
Base Шоу спочатку було озвучене непрофесійними акторами.
GRPO Шоу спочаткуfeatured аматорських акторів озвучення.

The trained model leaks the English source token *featured* into the Ukrainian output and concatenates it onto the previous word. Such fragments occur in $< 1\%$ of trained outputs but they are catastrophic for the affected sentences.

FLoRes+ uk→en, baseline vs. condition A. *Largest improvement* (judge $\Delta = +20$, $70 \rightarrow 90$):

Src Кожна станція має свою електронну адресу.
Ref Each station has its own electronic address.
Base Each station has its own e-mail address.
GRPO Each station has its electronic address.

The baseline renders the Ukrainian phrase for "electronic address" as "e-mail address"; the trained model preserves the literal meaning, matching context.

WMT24 uk→en, near-zero example. *Tied judge score* ($\Delta = 0$, $85 \rightarrow 85$):

Src Він побудував Wi-Fi дверний дзвінок.
Base He created a Wi-Fi doorbell.
GRPO He built a Wi-Fi doorbell.

Both translations are correct paraphrases. Cases like this account for the bulk of the 82–95% tied-pair fraction reported in Appendix C.