GRRM: Group Relative Reward Modeling for Machine Translation

Anonymous ACL submission

Abstract

While Group Relative Policy Optimization (GRPO) offers a powerful framework for LLM post-training, its effectiveness in open-ended domains like Machine Translation hinges on accurate intra-group ranking. We identify that standard Pointwise Quality Metrics (PQM) fall short in this context: candidates are evaluated in isolation, so the comparative context is missing for distinguishing fine-grained linguistic nuances. To address this, we introduce the Group Quality Metric (GQM) paradigm and its instantiation, the Group Relative Reward Model (GRRM). Unlike traditional independent scorers, GRRM jointly processes the entire candidate group, leveraging comparative analysis to rigorously resolve relative quality and adaptive granularity. Empirical evaluations confirm that GRRM achieves competitive ranking accuracy among all baselines; integrating GRRM into the GRPO training not only improves general translation quality but also unlocks reasoning capabilities comparable to state-of-the-art reasoning models.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has significantly advanced the reasoning processes of Large Language Models (LLMs), yielding remarkable performance in sophisticated domains such as mathematics and programming (Lambert et al., 2024; Guo et al., 2025a). Performing reinforcement learning on open-ended tasks, such as Machine Translation (MT), requires reward models to evaluate the responses generated by the policy model. Existing works predominantly employ Discriminative Reward Models (DRMs) based on the Bradley-Terry model (Cheng et al., 2025; Yang et al., 2025b). While DRMs have proven effective in improving translation quality in general domains, they have been criticized for lacking reasoning capabilities, which limits their effectiveness in challenging scenarios.

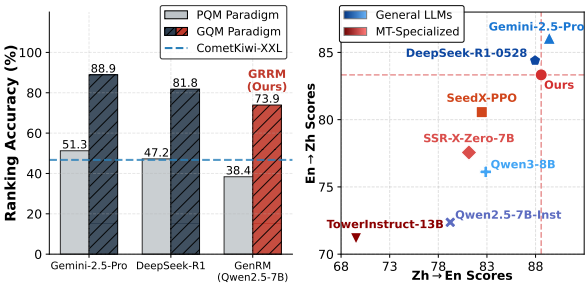

Figure 1: Performance on Seed-X-Challenge. **Left**: Ranking accuracy across paradigms. **Right**: Translation performance across General and MT-specialized LLMs.

Recently, the success of LLM-based generative evaluation (LLM-as-a-Judge) in MT (Kocmi and Federmann, 2023a,b) suggests the potential of Generative Reward Models (Zhang et al., 2025; Mahan et al., 2024, GenRMs). Ideally, the reasoning potential of GenRMs offers more accurate reward estimation, particularly for difficult samples. However, we identify a critical limitation when applying GenRMs within the widely used Group Relative Policy Optimization (Shao et al., 2024, GRPO) framework. Since GRPO estimates advantages based on the relative quality of responses within a group, the reward model must achieve accurate fine-grained intra-group ranking. Our investigation reveals that standard generative approaches, which evaluate responses independently (**Pointwise Quality Metric, PQM**), suffer from two fundamental flaws: the inability to distinguish subtle nuances and the vulnerability to misleading translations in challenging samples. These flaws lead to **score saturation**, where the model assigns identical maximal scores to candidates despite significant quality disparities. This lack of variance causes vanishing advantages in GRPO, depriving the policy of the gradient signals needed for optimization.

To address the limitations of the PQM paradigm, we propose the **Group Quality Metric (GQM)**. Specifically, GQM enables the model to process all candidate translations within a group collectively,

aiming to jointly estimate their quality ranking and assign relative scores. This cross-sample context provides consistent evaluation criteria, allowing the model to focus on distinguishing features between candidates. Furthermore, this paradigm allows the model to adaptively adjust its evaluation granularity according to the quality variance within the group.

However, current LLMs are weak when working in GQM style. We present the training pipeline of an instantiation of GQM, i.e. **Group Relative Reward Model (GRRM)**, a generative reward model structurally aligned with the group-relative nature of GRPO. Our GRRM is initialized from Qwen2.5-7B (Yang et al., 2024), cold-started via Supervised Fine-Tuning (SFT) on Chinese-English data, and subsequently optimized for ranking accuracy via RLVR. We demonstrate that both our trained GRRM and general LLMs evaluated under the GQM formulation consistently outperform their PQM counterparts, verifying the universality of the proposed metric. Notably, on challenging datasets, this paradigm delivers significant absolute gains of 30% to 40% in translation quality ranking accuracy compared to PQM (Fig. 1 Left). Furthermore, our GRRM exhibits strong cross-lingual generalization; despite being trained on a single language pair, it effectively supports multilingual translation optimization without additional adaptation.

We further validate our method in machine translation GRPO training, by using GRRM feedback to optimize the same Qwen2.5-7B backbone. For general domains, our model achieves an average 7.5-point improvement in BLEURT and a 16 point increase in LLM-judge scores over the SFT baseline across seven English-to-X (En2X) language pairs. Moreover, in challenging Chinese-English scenarios, our model performs comparably to DeepSeek-R1-0528 (Guo et al., 2025a) (Fig. 1 Right). Our analysis reveals that GRRM fosters the emergence of reasoning capabilities in translation models, which is crucial for solving complex translation challenges. In contrast, while DRMs can improve general translation quality, they are prone to reward hacking and lack the reasoning faculties required to guide models in challenging contexts.

2 Methodology

2.1 Ranking Sensitivity in GRPO

GRPO is a reinforcement learning paradigm that has proven highly effective for LLM post-training, particularly in complex reasoning domains.

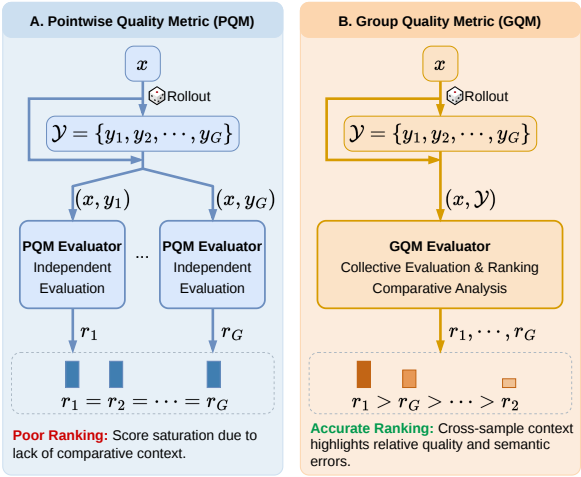

Figure 2: Comparison of PQM (A) and GQM (B).

Formally, for each query x, the policy π_θ samples a group of outputs $\mathcal{Y} = \{y_1, \ldots, y_G\}$. The policy is optimized to maximize the expected advantage of the generations. Specifically, the advantage A_i for the i-th candidate is derived by standardizing its reward r_i against the group distribution:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})} \quad (1)$$

where r_i is the scalar reward score assigned to y_i.

Eq. 1 underscores a critical property of GRPO: optimization is driven by the relative quality of candidates within a group, rather than their absolute scores. A candidate y_i is reinforced ($A_i > 0$) solely if it outperforms the group average. Consequently, GRPO depends on the reward model's ability to accurately rank candidates within $\mathcal{Y}$; any ranking inversion yields adversarial gradients that mislead policy optimization. This sensitivity motivates our investigation into the ranking limitations of current reward modeling paradigms.

2.2 Limitations of Pointwise Quality Metric

Current generative reward models typically function as a *Pointwise Quality Metric* (PQM), as illustrated in Figure 2 (A). Formally, PQM is defined as a mapping $S_{\text{PQM}} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ that assigns a scalar score r_i to a candidate y_i given an input x. Crucially, this evaluation is performed independently for each candidate. While this pointwise scoring is standard for general quality estimation, our analysis demonstrates that such independence fundamentally limits the model's ability to capture the fine-grained relative distinctions required for effective GRPO updates.

Inability to Distinguish Subtle Differences. LLMs frequently generate candidates that differ

only in fine-grained nuances. Without comparative context, PQM suffers from discriminative failure. As seen in Appendix Fig. 5 (Case 1), the model fails to perceive the stylistic superiority of one translation over another, treating them as equivalent.

Vulnerability to Misleading Candidates. More critically, independent evaluation leaves the reward model vulnerable to misleading candidates. Lacking a high-quality peer for contrast, even powerful models may overlook semantic discrepancies or localization errors, defaulting to surface fluency as proxies for quality. As shown in Appendix Fig. 5 (Case 2), PQM fails to penalize severe localization errors if they appear plausible in isolation.

Score Saturation. Both discriminative failure and susceptibility to misleading candidates culminate in a phenomenon we term **score saturation**. PQM frequently assigns identical maximal scores to candidates despite significant quality disparities, causing vanishing advantages in Eq. 1 and stalling policy optimization. A detailed empirical analysis is provided in Appendix A.

2.3 Group Quality Metric

To overcome the structural limitations of PQM, we introduce the *Group Quality Metric* (GQM), depicted in Figure 2 (B). Unlike the point-wise approach, GQM evaluates the policy-generated candidate group collectively. Formally, it maps a source x and a candidate set $\mathcal{Y}$ to a score vector: $S_{\text{GQM}} : \mathcal{X} \times \mathcal{Y}^G \to \mathbb{R}^G$. This holistic evaluation shifts the paradigm from absolute estimation to *relative ranking*.

By evaluating candidates jointly, GQM constructs a comparative context that circumvents the discriminative bottlenecks of PQM. This context mitigates score saturation, enabling the model to resolve fine-grained nuances that remain imperceptible under independent evaluation (Appendix Fig. 5, Case 1). Simultaneously, GQM introduces contrastive anchoring: plausible but erroneous translations (e.g., hallucinations) become salient when directly compared against high-quality peers (Appendix Fig. 5, Case 2). This mechanism ensures the reward signal captures both subtle stylistic preferences and critical semantic errors.

Crucially, GQM scores are inherently localized and do not map to a global absolute scale. This allows the model to adaptively adjust evaluation granularity: shifting focus to minor stylistic differences when group variance is low, while prioritizing major semantic discrepancies when significant quality divergences exist.

2.4 Group Relative Reward Model

While the GQM paradigm demonstrates efficacy when powered by frontier LLMs, directly integrating such massive models into the GRPO training loop is computationally intractable due to high latency and inference costs. We therefore introduce the Group Relative Reward Model (GRRM), an efficient instantiation of the GQM paradigm designed for high-throughput training loops.

GRRM is designed to process the source input x and the full candidate set $\mathcal{Y}$ jointly. For every tuple $(x, \mathcal{Y})$, the model is instructed to produce a structured output comprising: (1) a detailed comparative analysis of the candidates, (2) a predicted ranking, and (3) a set of scalar scores $\{r_1, \ldots, r_G\}$ that strictly adhere to the predicted ranking. This design enforces a Chain-of-Thought (CoT) reasoning process, ensuring numerical scores are grounded in explicit comparative analysis. Fig. 6 in Appendix shows an example of GRRM prompting and output.

We initialize GRRM via SFT on a curated dataset to establish instruction adherence and comparative logic. Subsequently, to sharpen the model's discriminative ability to distinguish subtle nuances, we employ RLVR to directly optimize its ranking accuracy. In this stage, we treat the GRRM itself as the policy to be optimized. To drive this optimization, we define a verifiable reward function, Ranking Accuracy, based on ground-truth orderings. Let $\mathbf{q} = \{q_1, \ldots, q_G\}$ be the ground-truth quality scores (or ranks) for the candidate group, and $\mathbf{r} = \{r_1, \ldots, r_G\}$ be the corresponding scores predicted by GRRM. We define the correctness of the pairwise relationship between any two candidates y_i and y_j as:

$$\mathbb{I}_{ij} = \mathbb{I}\left[\text{sgn}(\hat{r}_i - \hat{r}_j) = \text{sgn}(q_i - q_j)\right] \quad (2)$$

where $\text{sgn}(\cdot)$ is the sign function, and $\mathbb{I}(\cdot)$ is the indicator function. This formulation enforces strict alignment, requiring the predicted relationship—including ties—to match the ground truth. The final ranking accuracy reward R_{acc} is then calculated as the fraction of correctly ordered pairs over all distinct combinations in the group:

$$R_{\text{acc}} = \frac{1}{\binom{G}{2}} \sum_{1 \le i < j \le G} \mathbb{I}_{ij} \quad (3)$$

This metric functions as a normalized variant of the

Kendall's τ coefficient (Kendall, 1938), adapted to the range $[0, 1]$ with explicit handling of ties.

Additionally, to complement the sparse ranking signal and ensure the predicted scalar scores are well-distributed, we further use an auxiliary score consistency reward, detailed in Appendix B.

3 GRRM Evaluation

In this section, we provide a comprehensive evaluation of the proposed GRRM within the GQM framework, benchmarking the ranking accuracy (Eq. 3) against multiple baselines. Our experiments span three distinct settings: in-domain performance, multilingual generalization, and complex reasoning scenarios designed to stress-test reward modeling capabilities.

3.1 Experimental Setup

Datasets and Benchmarks. We construct our training data using the Chinese-English subset of TowerBlocks (Alves et al., 2024), comprising approximately 18.8k samples. We sample 2-4 translation candidates per source and annotate them with `Gemini-2.5-Pro` (Comanici et al., 2025) under the GQM paradigm. A held-out set of 512 samples is reserved as the model-annotated test set. For RLVR, we utilize the same underlying data as SFT.

We utilize Newstest2020 (Zh→En) with pSQM (professional Scalar Quality Metric) labels (Freitag et al., 2021) from the WMT20 metrics task. To evaluate cross-lingual generalization, we employ GeneralMT2022 with MQM (Multidimensional Quality Metric) annotations (Freitag et al., 2021) from WMT22. Beyond the in-domain Zh↔En pairs, this dataset includes English-to-German (En→De) and English-to-Russian (En→Ru) subsets. For these datasets, we retain up to 4 system outputs per source and derive ground-truth rankings based on the human-annotated scores.

To assess reasoning capabilities in complex linguistic scenarios, we construct a challenge set based on the Seed-X-Challenge test set (Cheng et al., 2025), which features idioms, slang, and domain-specific terminology in Zh↔En pairs. We design a binary ranking task to test the model's reasoning capability: for each source, we sample 4 translation candidates and select the best one according to BLEURT (Sellam et al., 2020). We then pair this chosen candidate with the expert human reference. Assuming the expert reference is strictly superior to the weak baseline's output, this setup requires the reward model to correctly identify the human translation as the winner.

Baselines. We include three categories of reward models for comparison. Unless otherwise noted, all trained models are initialized from `Qwen2.5-7B`.

- **LLM-as-a-Judge:** We test `Gemini-2.5-Pro` and `DeepSeek-R1-0528`. We evaluate them using both the standard PQM prompting and our proposed GQM prompting, explicitly instructing the models to use CoT reasoning.
- **DRMs:** We include CometKiwi-XXL (Rei et al., 2023), a leading reference-free Quality Estimation metric. Additionally, we train a BT-RM using the standard Bradley-Terry loss on pairwise preferences derived from our Gemini-annotated training data.
- **GenRMs:** In addition to our GRRM, we train a PQM-GenRM baseline using Gemini PQM annotations, following an RLVR optimization process similar to that of GRRM.

See Appendix D.1 for details on data construction and training hyperparameters.

3.2 Performance of GRRM

Table 1 presents the ranking accuracy of various reward models across held-out, general, and challenging datasets. Our key findings are as follows:

Dominance of GQM over PQM. GQM consistently outperforms the traditional PQM approach across all evaluator backbones. For LLM-as-a-Judge, employing GQM yields substantial improvements; for instance, `Gemini-2.5-Pro` achieves an average accuracy of 70.11% with GQM, compared to just 56.17% with PQM. Notably, this performance gap is maximized in the challenge scenario. While PQM performance for even state-of-the-art models degrades to near-random levels (e.g., 47.22% for DeepSeek-R1), GQM maintains robust accuracy. This trend holds for trained reward models as well, where GRRM (based on GQM) significantly outperforms PQM-GenRM. These results validate that independent evaluating (PQM) fails to capture fine-grained relative differences. By processing candidates collectively, GQM provides the model with a comparative context, enabling more robust and consistent ranking.

Effectiveness of GRRM and Reasoning. Our proposed GRRM achieves the highest average accuracy among all trained models. While we observe that RLVR optimization brings consistent

Model	Paradigm	held-out	NT20	GenMT22 (MQM)				Seed-X-Challenge	Avg.
		Zh↔En	Zh→En	En→Zh	Zh→En	En→De	En→Ru	Zh↔En	
Random	-	43.47	39.42	43.82	38.74	35.91	37.64	44.70	40.04
LLM-as-a-Judge (w/ Reasoning)									
Gemini-2.5-Pro	PQM	70.28	53.56	46.41	61.99	58.92	64.88	51.26	56.17
Gemini-2.5-Pro	GQM	-	**62.60**	**64.58**	**67.75**	**65.61**	**71.23**	**88.89**	**70.11**
DeepSeek-R1-0528	PQM	66.11	48.42	43.67	58.09	53.64	59.13	47.22	51.69
DeepSeek-R1-0528	GQM	**80.92**	61.98	64.38	65.79	63.26	69.15	81.82	67.73
Discriminative RMs (w/o Reasoning)									
CometKiwi-XXL	PQM	72.01	57.82	**66.49**	61.60	**61.20**	**67.12**	46.72	60.16
BT-RM	PQM	**82.62**	**58.16**	**66.49**	**64.92**	58.14	66.10	58.84	**62.11**
Generative RMs (w/ Reasoning)									
PQM-GenRM (SFT)	PQM	61.31	47.18	37.21	57.68	48.37	51.84	38.13	46.74
PQM-GenRM (RLVR)	PQM	64.25	49.21	39.42	60.37	49.16	54.82	38.38	48.56
GRRM (SFT)	GQM	79.75	56.77	59.45	64.29	58.65	64.80	69.95	62.32
GRRM (RLVR)	GQM	**83.19**	**57.57**	**62.12**	**65.55**	**60.75**	**66.39**	**73.93**	**64.38**

Table 1: Ranking accuracy (%) on the **held-out** test set (Gemini-annotated) versus human-annotated benchmarks and the challenge set. **Avg.** denotes the mean accuracy across all external benchmarks (excluding held-out). The best performance within each model category is **bolded**.

performance gains over the SFT cold-start, the magnitude of improvement is less significant than the paradigm shift from PQM to GQM. Crucially, the advantage of GRRM is most pronounced in the Seed-X Challenge set, achieving a ranking accuracy of 73.93%. This indicates that the reasoning capability is essential for verifying complex linguistic phenomena that statistical correlations alone cannot capture. Additionally, despite being trained solely on Zh-En data, GRRM exhibits strong cross-lingual generalization to unseen En→De and En→Ru pairs, performing on par with the CometKiwi-XXL baseline.

Limitations of Discriminative RMs. DRMs perform competitively on general domain benchmarks. This can be attributed to the Bradley-Terry loss, which implicitly and effectively models the ranking objective on preference pairs. However, their performance collapses on the challenge set due to the lack of generative reasoning capabilities, limiting their utility in guiding policies through complex optimization landscapes. For additional comparisons against discriminative baselines that incorporate limited contrastive context, see Appendix E.1.

4 Applying GRRM to MT Optimization

Building upon the validation of GRRM's ranking capabilities, we integrate it into the training loop to optimize the translation policy. This section evaluates whether GRRM-guided optimization improves MT performance and promotes the emergence of translation-oriented reasoning.

4.1 Experimental Setup

Training Pipeline. We adopt a two-stage strategy initialized with `Qwen2.5-7B`. First, we perform SFT to set up preliminary translation and reasoning skills using the same Chinese-English data described in Section 3 (with `Gemini-2.5-Pro` annotated CoT). Second, we optimize the model using GRPO with GRRM feedback based on the multilingual translation data from TowerBlocks, covering 10 languages with approximately 150k samples.

Since GRRM yields reference-free rewards, we implement Cross-Lingual Augmentation (CLA) by pairing source sentences with alternative target languages (e.g., Zh→De) from the dataset, while maintaining the total number of update steps constant.

Datasets. Our primary evaluation targets Zh↔En translation using WMT23 Zh→En (Kocmi et al., 2023), WMT24++ En→Zh (Deutsch et al., 2025), and the Seed-X-Challenge. For En→X, we report average performance across seven languages from WMT24++. Detailed results for WMT24++ En→X and WMT23 X→En are provided in Appendix E.3.

Evaluation Metrics. We employ a combination of standard metrics and LLM-based evaluation to provide a comprehensive assessment:

- **BLEURT:** We use BLEURT-20 as our primary automatic metric to measure semantic preservation against human references.
- **LLM-as-a-Judge:** Despite our findings in Section 3 regarding the superiority of GQM for ranking, we utilize the PQM paradigm for

system-level evaluation. We justify this choice based on two key factors: (1) **Reference Anchoring:** Our evaluation here includes human references. The reference serves as a strong anchor (conceptually similar to a group size of 2 in GQM), significantly stabilizing the judgment. (2) **System-Level Accuracy:** While PQM struggles with fine-grained intra-group ranking, it remains reliable for aggregating scores at the system level.

For Seed-X-Challenge, we further enhance the evaluation prompt by including the expert annotations provided with the dataset. These annotations highlight specific translation difficulties, enabling the judge to perform a more informed assessment alongside the reference.

- **Evaluator Models:** To balance overhead, we employ `DeepSeek-R1-0528` as the judge for all main results and `gpt-oss-120b` for ablation studies and supplementary results.

We also report additional reference-based results using MetricX-24 in Appendix E.2.

Baselines. We incorporate the following baselines:

- **General LLMs:** We test `Gemini-2.5-Pro` and `DeepSeek-R1-0528`, along with `Qwen3-8B` (Yang et al., 2025a) and `Qwen2.5-7B-Instruct` (Yang et al., 2024) to benchmark improvements against the base model family.
- **Translation-Specialized Models:** We compare against three state-of-the-art open models: (1) `TowerInstruct-13B` (Alves et al., 2024), a LLaMA2 (Touvron et al., 2023) derivative specifically adapted for translation tasks through continued pre-training and instruction tuning; (2) `SeedX-PPO` (Cheng et al., 2025), a 7B model pre-trained on high-quality multilingual datasets and subsequently enhanced via RL; and (3) `SSR-X-Zero-7B` (Yang et al., 2025c), a `Qwen2.5-7B` derivative which utilizes a self-rewarding RL framework specifically optimized for Zh↔En translation tasks.

Reasoning Configuration. For General LLMs, we explicitly prompt them to use CoT reasoning to maximize their potential. For specialized models, `SeedX-PPO` and `SSR-X-Zero-7B` leverage reasoning, while `TowerInstruct-13B` supports direct translation only.

4.2 Main Results

Table 2 presents the performance of our GRRM-optimized models compared to the SFT baseline and other state-of-the-art systems. The results indicate that integrating GRRM into the GRPO framework significantly enhances translation quality, particularly in scenarios requiring reasoning.

Compared to the `Qwen2.5-7B-SFT` baseline, our method achieves substantial improvements across all language pairs. In the general domain (WMT benchmarks), the model achieves an average gain of **+7.5 BLEURT** points and **+15.9 LLM-judge** points on En→X tasks. This successful optimization on multilingual data validates the strong cross-lingual generalization of GRRM, consistent with our metric analysis in Section 3. Furthermore, on Zh↔En tasks, our approach comprehensively surpasses all *Translation-Specialized Models* in LLM-judge scores, narrowing the gap with much larger proprietary models like `Gemini-2.5-Pro`.

The advantages of our approach are most pronounced in the challenge scenarios, i.e., Seed-X-Challenge. Our model equipped with Cross-Lingual Augmentation (CLA) achieves parity with the powerful reasoning model `DeepSeek-R1-0528`, even though the latter serves as our evaluator. Specifically, in the Zh→En subset, our model slightly outperforms `DeepSeek-R1-0528` (88.58 vs. 87.95), while remaining highly competitive in the En→Zh subset (83.33 vs. 84.40). We provide detailed case studies in Appendix E.5 to qualitatively demonstrate how our model successfully navigates these challenging samples with reasoning.

Our results also reveal a notable divergence between BLEURT and LLM-as-a-Judge scores. For instance, while `SeedX-PPO` achieves the highest BLEURT scores in several settings, its performance drops significantly under the scrutiny of the reasoning-based LLM judge. This discrepancy echoes our observations in Section 3, highlighting the limitations of traditional discriminative metrics like BLEURT when evaluating translation quality in challenging scenarios.

4.3 Ablation Study

We conduct ablation studies focusing on three key aspects: the choice of reward signals, the impact of training data distribution, and the role of reasoning in the policy model (see Table 3). We place extended analyses on the non-reasoning baseline and `Qwen2.5-7B-Instruct` in Appendix E.4.

Model	WMT Benchmarks						Seed-X-Challenge			
	Zh→En		En→Zh		En→X		Zh→En		En→Zh	
	BLEURT	R1-judge	BLEURT	R1-judge	BLEURT	R1-judge	BLEURT	R1-judge	BLEURT	R1-judge
General LLMs										
Gemini-2.5-Pro	**68.66**	**92.92**	**66.00**	**91.31**	**68.87**	**90.35**	**71.59**	**89.41**	**69.19**	**86.06**
DeepSeek-R1-0528	67.78	92.34	64.87	89.24	67.72	88.48	70.92	87.95	68.23	84.40
Qwen3-8B	63.03	89.72	57.58	84.15	57.25	77.37	61.17	82.88	57.78	76.12
Qwen2.5-7B-Instruct	67.31	88.49	59.92	80.51	58.72	72.51	66.59	79.23	62.75	72.37
Translation-Specialized Models										
TowerInstruct-13B	67.56	84.83	62.92	77.63	66.61	82.68	63.32	69.54	63.46	71.17
SeedX-PPO	**69.02**	90.47	**67.21**	87.98	**68.35**	**86.04**	69.37	82.47	**68.72**	80.56
SSR-X-Zero-7B	68.30	88.67	66.12	83.78	-	-	68.84	81.15	67.08	77.56
Qwen2.5-7B-SFT	67.07	87.78	59.99	76.98	57.14	67.91	67.65	80.91	62.36	72.42
+ GRPO (ours)	67.41	**92.24**	64.80	87.80	64.65	83.86	**69.55**	85.90	67.05	82.55
+ GRPO w/ CLA (ours)	67.39	92.09	63.91	**88.29**	64.50	83.71	69.25	**88.58**	67.07	**83.33**

Table 2: **MT performance on WMT and Seed-X-Challenge benchmarks.** We report BLEURT-20 and LLM-as-a-Judge scores evaluated by `DeepSeek-R1-0528`. The best performance within each category is highlighted in **bold**.

Ablation Configuration	WMT Benchmarks						Seed-X-Challenge			
	Zh→En		En→Zh		En→X		Zh→En		En→Zh	
	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge
Comparison of Reward Models										
SFT baseline (w/ Reasoning)	67.07	85.50	59.99	76.80	57.14	65.71	67.65	80.76	62.36	74.30
+ GRPO w/ BLEURT	**68.71**	88.21	**67.09**	85.52	**66.55**	82.00	69.69	83.35	**68.66**	81.85
+ GRPO w/ BT-RM	68.06	88.86	60.20	76.99	56.53	64.93	**69.74**	85.60	58.25	66.91
+ GRPO w/ PQM-GenRM	68.63	89.01	44.28	45.54	27.53	8.15	69.85	84.25	47.19	46.42
+ GRPO w/ GRRM	67.41	**89.85**	64.80	**87.91**	64.65	**83.86**	69.55	**87.30**	67.05	**84.36**
Cross-Lingual Generalization										
+ GRPO w/ GRRM (ZhEn)	66.53	89.11	63.07	86.88	61.70	79.18	69.26	86.16	66.20	83.86
Role of Reasoning										
SFT baseline (w/o Reasoning)	67.61	85.28	63.50	79.42	63.24	75.33	65.52	76.83	64.04	73.34
+ GRPO w/ GRRM	67.38	88.42	64.15	85.19	63.40	78.88	68.57	85.38	65.92	79.35

Table 3: **Ablation study on Reward Models and Reasoning configurations.** We compare different reward signals and training setups. We report BLEURT-20 and LLM-as-a-Judge scores evaluated by `gpt-oss-120b`.

Comparison of Reward Models. Optimizing directly against BLEURT yields the highest scores on the BLEURT metric itself, as expected. However, this gain does not translate effectively to LLM-judge scores, particularly on the Seed-X-Challenge (e.g., 81.85 vs. 84.36 on En→Zh). This suggests that discriminative metrics fail to guide the model through the complex reasoning processes required for challenging translations.

More critically, we observe severe reward hacking with BT-RM and PQM-GenRM. These models perform catastrophically on En→Zh and En→X tasks and suffer from off-target generation issues: apt to generate English responses for non-English targets. This indicates that standard discriminative or pointwise generative rewards struggle to distinguish instruction-following failures from translation quality issues. In contrast, GRRM provides robust feedback, effectively penalizing such deviations and preventing reward hacking.

Cross-Lingual Generalization We examine whether the cross-lingual generalization observed in our reward model (Section 3) transfers to Policy Optimization. We train a variant using only Chinese-English data, i.e., `GRPO w/ GRRM (ZhEn)`. While it improves performance on the Zh↔En pair, its performance on En→X lags significantly behind the model trained on multilingual data (79.18 vs. 83.86). This highlights a distinct difference between evaluation and generation: while the Reward Model demonstrates robust cross-lingual generalization, the Translation Policy exhibits significant transfer loss on unseen languages, indicating that explicit multilingual exposure remains crucial for optimal generation quality.

The Necessity of Reasoning. Finally, we assess the contribution of the reasoning process (CoT) to translation quality. It is evident that models without reasoning capabilities consistently underperform their reasoning counterparts. Even when optimized with GRRM, the non-reasoning model scores lower on both general benchmarks and challenge sets (e.g., 79.35 vs. 84.36 on Seed-X En→Zh). This validates the necessity of a "Reasoning × Reasoning" paradigm: the full potential of our framework is unlocked only when a reasoning-capable policy is paired with a reasoning-based reward model (GRRM), enabling the system to effectively plan, self-correct, and evaluate in complex scenarios.

5 Related Work

Reasoning for Machine Translation The integration of reasoning capabilities into Machine Translation has emerged as a promising direction (Chen et al., 2025; Liu et al., 2025a). Previous attempts to integrate reasoning into MT primarily relied on prompting or SFT (Feng et al., 2025b; Wang et al., 2025a; Cheng et al., 2025). However, Zebaze et al. (2025) observed that fine-tuning on synthetic CoT data alone often fails to outperform standard fine-tuning. Fundamentally, these SFT-based methods are not only constrained by the scarcity of high-quality supervision but also struggle to cultivate intrinsic reasoning capabilities.

Inspired by the success of reinforcement learning in mathematical and coding domains, recent works have adopted RL to encourage self-emergent reasoning. He et al. (2025) and Feng et al. (2025a) adapt the GRPO framework using discriminative metrics like Comet (Rei et al., 2020) and CometKiwi (Rei et al., 2023) to guide the model. Similarly, SSR-Zero (Yang et al., 2025c) employs the model itself to assign pointwise quality scores combined with Comet to estimate advantages, demonstrating that self-evaluation can drive policy improvement. By contrast, ExTrans (Wang et al., 2025b) leverages a superior large reasoning model to generate exemplar references and evaluates policy outputs by contrasting them against these anchors. However, these works employ reward mechanisms that are suboptimal for fostering robust reasoning. DRM-based methods lack the generative explanatory power to guide models through complex errors and are prone to reward hacking. The pointwise evaluation mechanism employed by SSR-Zero fails to capture the fine-grained relative rankings within a group.

Comparative and Reasoning-based Reward Modeling While machine translation evaluation has traditionally relied on PQM paradigm, employing pairwise comparative paradigms for judgment and reward estimation is not new in general open-ended domains (Li et al., 2024; Zhu et al., 2025). Ye et al. (2025) leverages self-generated contrastive pairs to train judges via DPO, enhancing robustness against bias compared to pointwise models. Liu et al. (2025b) introduces a pairwise judge with a knockout tournament for Best-of-N sampling, while Guo et al. (2025b) shows that CoT improves scoring accuracy by leveraging test-time compute.

However, these comparative approaches have yet to be effectively adapted to the group-wise context required by the GRPO framework. Existing pairwise methods are structurally misaligned with GRPO: estimating advantages for a group of responses via pairwise judges necessitates tournament-style evaluations (e.g., ELO ratings), which incur prohibitive computational costs. In contrast, our GRRM extends the comparative intuition to a list-wise context. By ranking all candidates in a single pass, our approach avoids iterative comparisons and empirically achieves 1.5× faster rewarding speeds than a standard pointwise generative model.

6 Conclusion

In this work, we identify and address the limitations of pointwise generative judges in GRPO training, specifically their insufficient sensitivity to distinguish intra-group quality differences in machine translation. To overcome this, we propose the Group Quality Metric (GQM), a paradigm that evaluates candidates jointly to capture fine-grained distinctions often missed by pointwise metrics. Building on this foundation, we introduce the Group Relative Reward Model (GRRM). Our experiments demonstrate that GRRM significantly outperforms existing discriminative and pointwise generative baselines, achieving robust ranking accuracy across diverse languages and challenging reasoning tasks. Crucially, integrating GRRM into the GRPO loop catalyzes the emergence of reasoning capabilities in translation models. This enables our approach to rival state-of-the-art systems like DeepSeek-R1 in complex scenarios while effectively mitigating reward hacking.

Limitations

We acknowledge two primary limitations of the proposed Group Quality Metric (GQM) and GRRM. First, unlike PQM which scales linearly, GQM is constrained by the maximum group size covered during training. Extrapolating to group sizes significantly beyond the training distribution may degrade ranking accuracy. However, this limitation does not hinder effective GRPO training. Recent work demonstrates that using small group sizes combined with larger batch sizes yields performance comparable to larger group configurations (Wu et al., 2025), rendering GQM's capacity sufficient for optimization.

Second, GQM relies on relative ranking, which introduces challenges in reward assignment when the overall group quality is low. Since GRRM treats intra-group ranking as the ground truth for advantage estimation, it may assign high relative scores to the "best" candidate even if all translations in the group are suboptimal. Consequently, GQM might fail to penalize the group globally as effectively as an absolute metric would. Integrating absolute quality constraints alongside relative ranking remains a promising direction for future work.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and 1 others. 2025. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *arXiv preprint arXiv:2502.11544*.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, and 1 others. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *arXiv preprint arXiv:2507.13618*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. *Preprint*, arXiv:2502.12404.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Jian Wu, and Zuozhu Liu. 2025a. MT-r1-zero: Advancing LLM-based machine translation via r1-zero-like reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18685–18702, Suzhou, China. Association for Computational Linguistics.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025b. TEaR: Improving LLM-based machine translation with systematic self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Preprint*, arXiv:2104.14478.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2025b. Reward reasoning model. *arXiv preprint arXiv:2505.14674*.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, and 1 others. 2025. R1-t1: Fully incentivizing translation capability in llms via reasoning learning. *arXiv preprint arXiv:2502.19735*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. Generative judge for evaluating alignment. In *ICLR*.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025a. New trends for modern machine translation with large reasoning models. *arXiv preprint arXiv:2503.10351*.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2025b. Pairjudge rm: Perform best-of-n sampling with knockout tournament. *arXiv preprint arXiv:2501.13007*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025c. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. Mt-ranker: Reference-free machine translation evaluation by inter-system ranking. In *The Twelfth International Conference on Learning Representations*.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, 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*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. DRT: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782, Vienna, Austria. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025b. Extrans: Multilingual deep reasoning translation via exemplar-enhanced reinforcement learning. *arXiv preprint arXiv:2505.12996*.

Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, and 1 others. 2025. It takes two: Your grpo is secretly dpo. *arXiv preprint arXiv:2510.00977*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Sen Yang, Yu Bao, Yu Lu, Jiajun Chen, Shujian Huang, and Shanbo Cheng. 2025b. EnAnchored-X2X: English-anchored optimization for many-to-many translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21304–21317, Suzhou, China. Association for Computational Linguistics.

Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, and Sitong Wang. 2025c. Ssr-zero: Simple self-rewarding reinforcement learning for machine translation. *arXiv preprint arXiv:2505.16637*.

Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. 2025. Learning llm-as-a-judge for preference alignment. In

The Thirteenth International Conference on Learning Representations.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Armel Zebaze, Rachel Bawden, and Benoît Sagot. 2025. Llm reasoning for machine translation: Synthetic data generation over thinking tokens. *arXiv preprint arXiv:2510.11919*.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025. Generative verifiers: Reward modeling as next-token prediction. In *The Thirteenth International Conference on Learning Representations*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. COMET-poly: Machine translation metric grounded in other candidates. In *Proceedings of the Tenth Conference on Machine Translation*, pages 887–904, Suzhou, China. Association for Computational Linguistics.

A Empirical Analysis of Score Saturation

As discussed in Section 2.2, when a model is used under the standard Pointwise Quality Metric (PQM) paradigm, it often produces overly optimistic scores and, in particular, frequently assigns the maximum possible score. We refer to this phenomenon as **score saturation**. This reduces reward variance within a candidate group, shrinking the standardized advantages and erasing the learning signal.

In this section, we present an empirical analysis to substantiate our claim that PQM suffers from severe score saturation compared with GQM, which in turn impedes effective policy optimization.

A.1 Setup

Our datasets follow Section 3.1: the held-out Zh↔En set, NT20 (Zh→En), GeneralMT22 (En→Zh, Zh→En, En→De, En→Ru), and Seed-X-Challenge (Zh↔En).

Saturation in LLM-based evaluation. For each dataset, we compute (i) the average score and (ii) the saturation rate, i.e., the fraction of candidates receiving the *maximum allowed score* under the corresponding prompting protocol. We report results for two advanced LLM judges (i.e., `Gemini-2.5-Pro` and `DeepSeek-R1-0528`) under both PQM and GQM prompting.

Saturation during GRPO optimization. To connect saturation to optimization dynamics, we also track the reward values observed during GRPO training of the translation policy when using PQM-GenRM versus our GRRM as the reward provider.

A.2 Results

PQM yields systematic saturation. As shown in Figure 3, across all seven datasets and for both judges, PQM consistently produces higher average scores than GQM, accompanied by markedly higher saturation rates. For example, with `Gemini-2.5-Pro`, PQM saturation ranges from 28.24% to 64.02%, while GQM reduces it to 22.04%–39.63%. With `DeepSeek-R1-0528`, saturation drops even more sharply—from 18.58%–47.41% (PQM) to 3.41%–13.12% (GQM). This indicates that pointwise PQM prompting tends to over-assign the maximum score, whereas GQM's joint comparison makes differences among candidates more salient and alleviates saturation.

PQM-based rewards saturate early during GRPO. Figure 4 further shows that PQM-based reward modeling can lead to an early reward plateau during policy optimization. PQM-GenRM rewards increase rapidly and approach a near-ceiling regime early in training, suggesting that the policy quickly learns to trigger near-maximal scores from the PQM reward model (i.e., reward hacking and saturation), after which the reward provides limited additional shaping.

In contrast, GRRM rewards increase more gradually and continue to fluctuate in a non-saturated regime. This pattern is consistent with GRRM providing a more discriminative and robust signal that remains informative throughout optimization, aligning with GRPO's reliance on within-group comparisons.

B Auxiliary Score Consistency Reward

To encourage the model to generate calibrated numerical scores that reflect the magnitude of quality

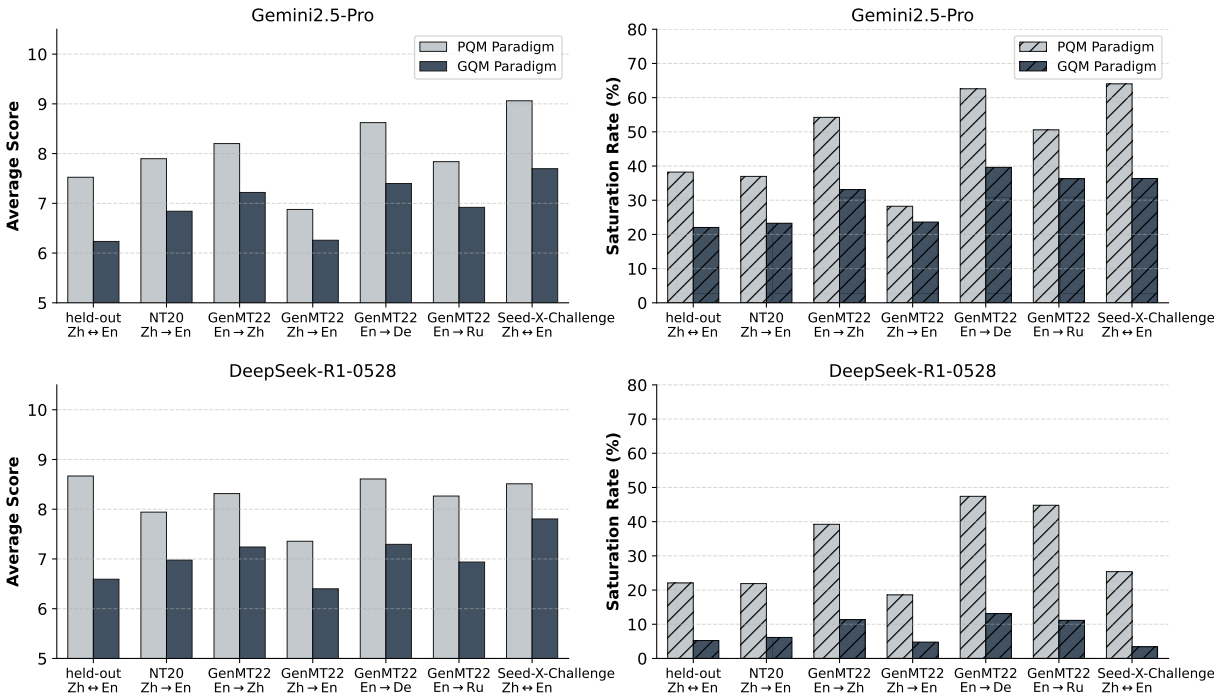

Figure 3: **Analysis of score saturation across seven benchmarks.** We compare the **Average Score** (left) and **Saturation Rate** (right) under PQM and GQM paradigms using `Gemini-2.5-Pro` and `DeepSeek-R1-0528`. PQM consistently exhibits score inflation and high saturation, whereas GQM effectively mitigates this issue, providing more discriminative evaluation.

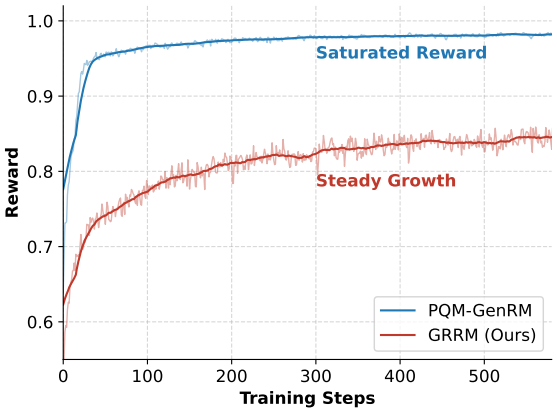

Figure 4: **Reward trends during GRPO training with 30-step moving averages, using PQM-GenRM vs. GRRM as reward providers (scores normalized to $[0, 1]$).** The PQM-GenRM rewards saturate early whereas GRRM remains in a non-saturated regime and continues to provide discriminative training signal.

differences, we augment the standard ranking accuracy reward with a score consistency objective. The GRRM is instructed to output a structured response concluding with two specific components: an explicit ranking string (e.g., $A > B = C$) and a dictionary of scalar scores (e.g., $\{A : 6, B : 5, C : 5\}$).

Internal Consistency Constraint. Before evaluating accuracy, we enforce a strict self-consistency check to ground the numerical outputs. Let $\hat{\tau}_{\text{text}}$ be the explicit ranking string generated by the model, and $\hat{\tau}_{\text{score}}$ be the ranking induced by sorting the generated scalar scores $\hat{\mathbf{s}}$. We define a binary consistency gate C_{gate}:

$$C_{\text{gate}} = \mathbb{I}(\hat{\tau}_{\text{text}} \equiv \hat{\tau}_{\text{score}}) \tag{4}$$

This gate is applied in the reward calculation to penalize inconsistency.

Margin-Aware Score Reward. For samples passing the consistency gate, we calculate the Score Consistency Reward (R_{score}). Unlike simple regression losses (e.g., MSE), which can be sensitive to the absolute scale of ground-truth labels, our metric focuses on preserving the *relative margins* between candidates. Let q_i, q_j be the ground-truth scores and $\hat{s}_i, \hat{s}_j$ be the predicted scores for candidates i and j. We compute the absolute margin error δ_{ij} for every pair:

$$\delta_{ij} = |(\hat{s}_i - \hat{s}_j) - (q_i - q_j)| \tag{5}$$

We then map this error to a reward value using a discrete kernel function $K(\cdot)$, designed to penalize

deviations while allowing for minor integer fluctuations. Based on our empirical tuning, we define:

$$K(\delta) = \begin{cases} 1.0 & \text{if } \delta = 0 \\ 0.6 & \text{if } \delta = 1 \\ 0.2 & \text{if } \delta = 2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The final score consistency reward is the average kernel value across all pairs:

$$R_{\text{score}} = \frac{1}{\binom{G}{2}} \sum_{1 \le i < j \le G} K(\delta_{ij}) \tag{7}$$

Total Reward. The final reward used for RLVR optimization is the sum of the ranking accuracy (Eq. 3) and the score consistency reward, gated by the internal consistency check:

$$R_{\text{total}} = C_{\text{gate}} \cdot (R_{\text{acc}} + R_{\text{score}}) \tag{8}$$

This formulation ensures that the model is optimized to produce reasoning that results in both correct ordering and precise, margin-aware numerical estimations.

PQM-GenRM Configuration. For the **PQM-GenRM** baseline, which evaluates candidates independently, we adapt the reward to target absolute accuracy. Unlike the margin-based approach in GRRM, we compute the absolute error $\delta = |\hat{s} - q|$ between the predicted score $\hat{s}$ and the ground truth q. The final reward utilizes a similar discrete kernel function $K(\delta)$ defined above to penalize deviations, ensuring a fair comparison between the optimization objectives.

C Stabilizing Group Relative Policy Optimization

Standard GRPO often exhibits training instability, particularly in the mid-training phase, where the model can become overly sensitive to negative rewards. This frequently leads to sharp distribution shifts and even collapse. To mitigate this, we propose **Stable GRPO**, which incorporates an auxiliary objective to anchor the policy updates to high-quality trajectories.

Recall that GRPO optimizes the policy by maximizing a clipped surrogate objective based on group relative advantages[1]:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \mathcal{L}_{i,t}^{\text{CLIP}}(\theta) \right], \tag{9}$$

[1]We omit the KL regularization term for brevity.

where the clipping objective is:

$$\mathcal{L}_{i,t}^{\text{CLIP}}(\theta) = \min \Big(w_{i,t}(\theta) A_i,\ \text{clip}\big(w_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\big) A_i \Big), \tag{10}$$

and the importance ratio is:

$$w_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}. \tag{11}$$

Building upon the standard GRPO objective, we introduce an advantage-weighted SFT term:

$$\mathcal{J}_{\text{StableGRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \Big(\mathcal{L}_{i,t}^{\text{CLIP}}(\theta) + \gamma \mathcal{L}_{i,t}^{\text{SFT}}(\theta) \Big) \right] \tag{12}$$

Here, $\mathcal{L}_{i,t}^{\text{SFT}}(\theta)$ applies a standard log-likelihood maximization, strictly gated by positive advantages:

$$\mathcal{L}_{i,t}^{\text{SFT}}(\theta) = \max(0, A_i) \ln \pi_\theta(y_{i,t} | x, y_{i,<t}) \tag{13}$$

where γ controls the strength of this regularization.

The key design is that $\mathcal{L}^{\text{SFT}}$ is applied only to samples with positive advantages ($A_i > 0$), complementing $\mathcal{L}^{\text{CLIP}}$ by strengthening updates on positive samples in the non-clipped regime and serving as an unclipped fallback signal when clipping suppresses the surrogate gradient.

We apply Stable GRPO for machine translation optimization only. For reasoning-based translation models, we set $\gamma = 0.1$ (increasing to $\gamma = 0.2$ for the Cross-Lingual Augmentation variant). For non-reasoning translation models, we set $\gamma = 1.0$.

D Implementation Details

In this section, we detail the training configurations for both the Reward Model (GRRM) and the Machine Translation Policy Optimization.

D.1 Reward Model Implementation

Training Data Construction. We construct the training dataset using the Chinese-English subset of TowerBlocks (MT), by sampling group-level candidates using a seed translation model. Specifically, we employ the `Qwen2.5-7B` model fine-tuned on the TowerBlocks (MT) dataset (approximately 150k samples). This seed model is also used as SFT baseline (w/o Reasoning) in Section 4 and Appendix E.4. We apply standard sampling with a temperature of $T = 1.0$. For each source sentence, we sample a group of N candidates, where $N \in \{2, 3, 4\}$ with a frequency ratio of 1:1:3. To

enhance candidate diversity and ensure the presence of high-quality samples within the generated groups, we inject the ground-truth reference into the candidate pool with a probability of 0.5. These constructed groups are subsequently annotated by `Gemini-2.5-Pro`.

Evaluation Data Processing. For the human-annotated test sets (Newstest2020 and GeneralMT2022), we calculate the final quality score for each translation candidate by averaging the ratings across raters. From the available system outputs for each source, we retain a subset of 2 to 4 candidates, following the same distribution ratio as the training data. However, we observe that randomly selected candidates often possess similar human quality scores (clustering near the ceiling), which introduces ambiguity into the ranking evaluation. Therefore, to ensure a rigorous assessment of the model's ability to distinguish quality variances, we enforce a filtering constraint: the constructed group must include the candidates with the minimum and maximum human scores for that entry.

Training Hyperparameters. We employ 16 Nvidia A100 (80GB) GPUs for all training experiments detailed in this paper, including both reward modeling and the subsequent machine translation policy optimization.

In the SFT stage, we optimize the model for 3 epochs with a global batch size of 64. We employ a cosine learning rate scheduler with a peak learning rate of 6×10^{-6} and a warmup ratio of 0.1.

In the RLVR stage, we adopt Group Sequence Policy Optimization (Zheng et al., 2025, GSPO), an enhanced variant of GRPO that uses sequence-level importance ratios to improve training stability and efficiency.

We set the learning rate to 1×10^{-5} with a cosine scheduler, decaying to a minimum ratio of 0.2. The training is conducted with a total batch size of 512 and a PPO mini-batch size of 128. For each prompt, we generate $G = 8$ rollouts with a maximum generation length of 4096 tokens. The model is trained for a single epoch. Following recent practices in reasoning model training, we disable the KL divergence penalty to encourage broader exploration of reasoning paths. Additionally, following Dr. GRPO (Liu et al., 2025c), we remove the standard deviation normalization terms from the advantage calculation in GRPO.

D.2 MT Optimization Implementation

SFT Hyperparameters. For the policy model cold-start, we perform Supervised Fine-Tuning for 1 epoch with a global batch size of 64. We utilize a cosine learning rate scheduler, setting the peak learning rate to 1×10^{-5} with a warmup ratio of 0.1. For the reasoning baseline, SFT uses the Zh-En data described in Section 3. For the non-reasoning baseline used in the ablation experiments, we instead use the full multilingual TowerBlocks (MT) dataset to avoid the severe Zh-En overfitting observed under restricted non-reasoning SFT, where the model failed to generalize to other languages during RL.

GRPO Hyperparameters. We utilize GSPO and apply the same stabilization enhancement in Appendix C on top of GSPO for MT policy optimization. The configuration largely mirrors the GRRM RLVR stage (Appendix D.1), with specific adjustments to suit the translation task. We switch to a constant learning rate scheduler to maintain stable updates throughout the training process. The number of rollouts per prompt is set to $G = 4$, aligning with the group size capacity of our reward model.

To enhance optimization stability and reduce the gradient norm, we scale the rewards to the range $[0, 0.1]$. Furthermore, following DAPO (Yu et al., 2025), we employ a soft length penalty with the overlong buffer set to 2048 tokens.

To ensure fair comparison across different data settings, we adjust the number of training epochs to keep the total number of optimization steps consistent. Specifically, we train for 2 epochs on the standard multilingual TowerBlocks (MT) dataset (150k samples). For the CLA variant, we train for 1 epoch, while for the Chinese-English only subset, we extend training to 8 epochs.

E Extended Analyses

E.1 Additional Contrastive Discriminative Baselines

In addition to the main baselines discussed in Section 3, we further compare GRRM with two recent MT evaluation models that introduce candidate-level contrastive context while still producing discriminative scores: MT-Ranker (Moosa et al., 2024) and COMET-poly (Züfle et al., 2025). This comparison is designed to disentangle the effects of two key components in our framework: (1) access to contrastive context among translation candidates,

Case 1: Fine-grained Lexical Nuance

Source: 联合国称阿富汗民主取决于女性的政治地位

Candidate Translations:

A: United Nations says Afghanistan's democracy depends on women's **political status**.
B: U.N. says Afghanistan democracy depends on women's **political role**.
C: Afghan democracy depends on women's **political clout**, U.N. says.

PQM Evaluation (Independent)

Ranking: A = B = C **Scores:** A:10, B:10, C:10
Analysis: PQM suffers from score collapse. It fails to distinguish the precision of "status" (A) versus the slightly broader "role" (B) or the informal register of "clout" (C), assigning maximum scores to all.

GQM Evaluation (Group Relative)

Ranking: A > B > C **Scores:** A:10, B:9, C:8
Analysis: GQM successfully captures fine-grained distinctions. It identifies "status" as the most precise equivalent for the source term, while penalizing "clout" for being stylistically inconsistent with the UN subject matter.

Case 2: Localization & Logical Error

Source: 今天所有餐品七五折 *(Lit: "Seven-five discount" → 75% of price → 25% off)*

Candidate Translations:

A: All menu items are **75% off** today.
B: All menu items are **25% off** today.

PQM Evaluation (Independent)

Ranking: A = B **Scores:** A:10, B:10
Analysis: Vulnerable to misleading fluency. "75% off" is grammatically perfect, so PQM fails to detect the localization error where "7.5 Zhe" means 25% off, not 75%.

GQM Evaluation (Group Relative)

Ranking: B > A **Scores:** A:1, B:10
Analysis: By contrasting the two options, the numerical discrepancy becomes salient. GQM correctly identifies that A contains a critical localization error despite its superficial fluency.

Case 3: Hallucination vs. Stylistic Stiffness

Source: While most output is human-readable, bar code printers are an example of an expanded use for printers.

Candidate Translations:

A: 大多数打印机的输出是可读的，而且，条形码打印机可以输出条形码，接触到更广泛的应用范围。
(Over-translation: "output barcodes"; Unidiomatic: "contact... range")
B: 然而，虽然最普遍的打印输出是可读的人类语言，但条形码打印机的示例显示打印机能用于其他用途。
(Semantically correct but stylistically stiff/repetitive)

PQM Evaluation (Independent)

Ranking: A > B **Scores:** A:5, B:4
Analysis: PQM fails to detect the unidiomatic usage ("接触到") and hallucination in A, likely over-penalizing B for stiffness.

GQM Evaluation (Group Relative)

Ranking: B > A **Scores:** A:2, B:8
Analysis: GQM prioritizes semantic accuracy. Comparing the two, it identifies the hallucinated content in A as a severe flaw, ranking the accurate B higher.

Case 4: Identifying Semantic Errors

Source: Garland is the chief judge for the U.S. Court of Appeals for the District of Columbia **Circuit**.

Candidate Translations:

A: 加兰是美国哥伦比亚特区上诉法院首席法官。
(Misses "Circuit" / Federal distinction)
B: 加兰德是美国哥伦比亚特区联邦巡回上诉法院的首席法官。
(Correctly translates specific court name)

PQM Evaluation (Independent)

Ranking: A = B **Scores:** A:10, B:10
Analysis: PQM overlooks the omission of "Circuit" in A. The sentence is fluent and plausible in isolation, so the semantic error is missed.

GQM Evaluation (Group Relative)

Ranking: B > A **Scores:** A:6, B:10
Analysis: In the group context, the contrast makes the omission of "Circuit" in A salient, penalizing the semantic error.

Figure 5: Comparison of Pointwise Quality Metric (PQM) and Group Quality Metric (GQM) across four distinct scenarios. **Case 1 & 2 (Top):** Demonstrate GQM's ability to resolve fine-grained stylistic nuances and identify localization errors that PQM misses due to independent evaluation. **Case 3 & 4 (Bottom):** Illustrate how GQM uses contrastive context to detect hallucinations and semantic omissions that appear fluent in isolation. Case 1,3,4 were conducted using `Gemini-2.5-Pro` and Case 2 was conducted using `DeepSeek-R1-0528`.

and (2) generative reasoning over the full candidate group.

Since these methods are not natively designed for our multi-candidate group ranking setting, we adapt them as follows. MT-Ranker is a pairwise judge; for datasets with more than two candidates per source, we aggregate pairwise outcomes into a global ordering using the ELO rating procedure. For Seed-X-Challenge, which is formulated as a binary ranking task, MT-Ranker is evaluated directly on pairwise comparisons. COMET-poly can take additional candidates as contrastive context, but it only assigns a score to the first candidate rather than producing a native list-wise ranking. We use the `COMET-poly-cand1-wmt25` model and randomly select another candidate from the same group as its contrastive context.

Table 4 shows that contrastive discriminative evaluators outperform purely pointwise metrics in challenging two-candidate settings, but still lag behind GRRM. On Seed-X-Challenge, where neither ELO aggregation nor group-size mismatch is involved, MT-Ranker-XXL and COMET-poly reach 65.66% and 66.60% ranking accuracy, respectively—substantially above CometKiwi-XXL and BT-RM, confirming the value of candidate-level contrastive context for detecting subtle translation errors. However, both

Model	held-out	NT20	GenMT22 (MQM)				Seed-X-Challenge	Avg.
	Zh↔En	Zh→En	En→Zh	Zh→En	En→De	En→Ru	Zh↔En	
Additional MT evaluation baselines								
MT-Ranker-XXL	65.82	51.37	56.59	56.22	50.09	55.96	65.66	55.98
COMET-poly	54.13	46.49	54.78	49.37	49.64	54.71	66.60	53.60
Reference rows								
CometKiwi-XXL	72.01	57.82	**66.49**	61.60	**61.20**	**67.12**	46.72	60.16
BT-RM	82.62	**58.16**	**66.49**	64.92	58.14	66.10	58.84	62.11
GRRM (RLVR)	**83.19**	57.57	62.12	**65.55**	60.75	66.39	**73.93**	**64.38**

Table 4: Additional ranking accuracy (%) with MT-Ranker and COMET-poly, compared with representative baselines from Table 1. **Avg.** denotes the mean accuracy across external benchmarks, excluding the held-out set.

remain well below GRRM at 73.93%, suggesting that contrastive context alone is not enough; generative reasoning over candidate differences provides additional gains on idioms, slang, terminology, and other difficult phenomena.

A similar pattern holds on the multi-candidate benchmarks. MT-Ranker-XXL and COMET-poly achieve average accuracies of 55.98% and 53.60%, respectively, compared with 64.38% for GRRM. This drop reflects the limitation of local contrastive scoring without a global view of the full candidate set. ELO aggregation converts pairwise preferences into a group ranking, but cannot fully recover listwise comparative information. Likewise, COMET-poly uses other candidates only as auxiliary context when scoring a single candidate, rather than jointly reasoning over the entire group.

Overall, the results suggest two conclusions: contrastive context is useful for translation reward modeling, as evidenced by the gains of MT-Ranker and COMET-poly on Seed-X-Challenge; but the full GQM formulation in GRRM is more effective, especially when GRPO requires accurate intra-group ranking over multiple policy samples. The consistent gap further suggests that explicit reasoning is crucial: it not only exposes candidate differences, but also helps the model make more reliable decisions on subtle and linguistically complex errors.

E.2 Additional Results with MetricX-24

To complement the main results reported with BLEURT and LLM-as-a-Judge, we further evaluate all systems using **MetricX-24** in reference-based mode. Specifically, we use the `MetricX-24-Hybrid-XXL` evaluator in the `bfloat16` variant, producing regression-style quality scores clipped to the range $[0, 25]$, where lower values indicate better translation quality.

As shown in Table 5, the overall conclusions remain consistent with the main results. Among general LLMs, `Gemini-2.5-Pro` achieves the strongest MetricX-24 performance. Among translation-specialized models, our GRPO-based variants substantially improve over the SFT baseline across all evaluation sets, and `+GRPO w/ CLA` obtains the best or tied-best results on most settings within our model family. These results provide an additional validation of the gains brought by GRRM-based policy optimization.

E.3 Full Multilingual Benchmarks

We provide the detailed breakdown of the aggregated performance reported in the main results. Table 6 details the En→X performance across seven European and Slavic languages from WMT24++, while Table 7 presents the WMT23 X→En results for German, Japanese, and Russian.

The results demonstrate that our GRPO-based optimization yields consistent improvements over the SFT baseline across diverse language families. For WMT23 X→En benchmark, while `SeedX-PPO` generally leads in BLEURT, our method achieves superior or competitive LLM judge scores, particularly in directions like Ja→En and Ru→En.

E.4 Extended Ablation Studies

In this section, we provide supplementary experiments to further validate the robustness and universality of our proposed framework. Table 8 presents the detailed results.

Reward Model Stability on Non-Reasoning Baselines. We first investigate whether the instability of certain reward models (observed in Section 4.3) persists when optimizing a standard, non-reasoning translation policy. As shown in the first block of Table 8, the **PQM-GenRM** continues to exhibit severe failure modes, particularly on En→X

Model	WMT Benchmarks			Seed-X-Challenge	
	Zh→En	En→Zh	En→X	Zh→En	En→Zh
General LLMs					
Gemini-2.5-Pro	**1.54**	**2.11**	**2.42**	**1.99**	**1.98**
DeepSeek-R1-0528	1.70	2.27	2.77	2.32	2.12
Qwen3-8B	3.47	4.68	5.30	4.38	5.13
Qwen2.5-7B-Instruct	2.13	3.75	5.43	3.44	3.54
Translation-Specialized Models					
TowerInstruct-13B	2.16	3.18	3.23	4.50	3.01
SeedX-PPO	**1.63**	**1.98**	**2.95**	**2.51**	**1.80**
SSR-X-Zero-7B	1.80	2.43	-	2.90	2.48
Qwen2.5-7B-SFT	2.05	4.09	6.02	3.19	4.58
+ GRPO (ours)	1.68	2.37	3.41	**2.56**	2.37
+ GRPO w/ CLA (ours)	**1.67**	2.47	3.36	**2.56**	2.26

Table 5: **MetricX-24 results on WMT and Seed-X-Challenge benchmarks.** Lower is better (↓). The best performance within each category is highlighted in **bold**.

Model	WMT24++ En→X Detailed Results													
	En→De		En→Es		En→Fr		En→It		En→Nl		En→Pt		En→Ru	
	BLEURT	R1-judge	BLEURT	R1-judge	BLEURT	R1-judge	BLEURT	R1-judge	BLEURT	R1-judge	BLEURT	R1-judge	BLEURT	R1-judge
General LLMs														
Gemini-2.5-Pro	**71.75**	**89.79**	**71.82**	**91.10**	**64.63**	**91.03**	**71.33**	**91.14**	**72.21**	**89.54**	**63.41**	**90.00**	**66.91**	**89.88**
DeepSeek-R1-0528	69.61	88.18	71.17	90.68	63.53	89.44	70.62	89.57	71.26	87.10	62.91	87.61	64.90	86.76
Qwen3-8B	63.90	77.13	64.58	82.71	40.45	76.21	62.53	78.98	62.89	71.37	56.80	80.58	49.62	74.62
Qwen2.5-7B-Instruct	59.43	68.94	62.84	78.61	54.47	76.80	59.84	71.83	61.44	64.21	56.64	76.82	56.41	70.39
Translation-Specialized Models														
TowerInstruct-13B	69.11	82.34	68.68	84.26	61.51	83.33	70.35	84.40	70.49	82.09	63.37	83.30	62.73	79.02
SeedX-PPO	**71.15**	**86.03**	**70.63**	**87.51**	**62.76**	**86.31**	**71.19**	**86.69**	**71.69**	**84.20**	**65.25**	85.95	**65.76**	**85.62**
Qwen2.5-7B-SFT	59.06	65.68	61.91	74.25	51.33	71.41	60.12	66.59	58.10	60.05	55.77	72.98	53.72	64.44
+ GRPO (ours)	68.12	82.63	66.59	86.91	59.78	85.27	67.43	84.07	67.04	79.30	61.24	85.65	62.32	83.22
+ GRPO w/ CLA (ours)	67.92	82.60	66.44	86.76	60.18	85.84	66.98	83.67	66.50	78.62	61.12	**86.43**	62.39	82.07

Table 6: **Detailed breakdown of WMT24++ En→X results.** We report BLEURT-20 and LLM-as-a-Judge scores evaluated by `DeepSeek-R1-0528`.

tasks (dropping to a score of 7.13), confirming that the vulnerability to reward hacking is inherent to the reward model itself rather than specific to reasoning policies. In contrast, our **GRRM** consistently outperforms other reward signals (BLEURT, BT-RM) across most benchmarks. However, it is worth noting that even with the best reward model (GRRM), the non-reasoning baseline (Score 79.35 on Seed-X En→Zh) still significantly lags behind the reasoning-enhanced policy reported in the main text (Score 84.36). This further corroborates our conclusion that the "Reasoning × Reasoning" paradigm is essential for peak performance.

Generalization to Instruction-Tuned Models. To assess the universality of our approach, we apply our GRPO training with GRRM directly to a standard instruction-tuned model, `Qwen2.5-7B-Instruct`, without any task-specific SFT warm-up. As shown in the second block of Table 8, our method yields substantial improvements over the strong base model. Notably, on the unseen En→X task, the performance improves by over 11 points on the `oss-judge` metric (70.84 → 82.44). This result demonstrates that our framework is not limited to specialized translation models but can serve as a general-purpose alignment technique to enhance the multilingual capabilities of off-the-shelf LLMs.

E.5 Case Studies: Emergence of Reasoning

To provide a qualitative perspective on the quantitative gains observed in the Seed-X-Challenge, Figure 7 illustrates the emergence of reasoning capabilities in our GRRM-optimized model. Unlike standard SFT models that often rely on direct surface-level mapping, our model demonstrates a "think-before-translating" mechanism that is crucial for resolving cultural ambiguities.

In the scenario of **Translating Idiomatic Expressions** (Case 1), the model successfully navigates the English proverb "The grass is always greener on the other side." Instead of producing a rigid or literal translation (e.g., regarding the color

Model	WMT23 X→En Detailed Results							
	De→En		Ja→En		Ru→En		Average	
	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge
General LLMs								
Qwen3-8B	69.12	**88.65**	60.87	**80.11**	65.42	**84.66**	65.14	**84.47**
Qwen2.5-7B-Instruct	**71.75**	86.21	**66.23**	79.03	**71.04**	83.80	**69.67**	83.01
Translation-Specialized Models								
TowerInstruct-13B	74.48	90.04	66.74	75.66	72.98	86.46	71.40	84.06
SeedX-PPO	**75.43**	**92.05**	**70.19**	84.99	**74.40**	89.82	**73.34**	**88.95**
Qwen2.5-7B-SFT	71.51	85.54	66.38	78.08	71.15	84.15	69.68	82.59
+ GRPO (ours)	70.25	90.69	67.72	**85.50**	71.93	**90.14**	69.97	88.78
+ GRPO w/ CLA (ours)	70.83	90.47	67.73	84.98	71.83	89.25	70.13	88.23

Table 7: **Detailed breakdown of WMT23 X→En results.** We report BLEURT-20 and LLM-as-a-Judge scores evaluated by `gpt-oss-120b`.

Supplementary Configuration	WMT Benchmarks						Seed-X-Challenge			
	Zh→En		En→Zh		En→X		Zh→En		En→Zh	
	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge
Comparison of Reward Models for Non-Reasoning Baselines										
SFT baseline (w/o Reasoning)	67.61	85.28	63.50	79.42	63.24	75.33	65.52	76.83	64.04	73.34
+ GRPO w/ BLEURT	68.36	86.46	65.28	81.19	65.12	77.49	67.21	79.17	65.25	75.99
+ GRPO w/ BT-RM	68.26	88.35	64.67	85.44	63.33	78.21	69.41	82.88	66.00	78.85
+ GRPO w/ PQM-GenRM	68.02	87.55	38.88	36.08	27.14	7.13	68.32	82.10	42.49	39.85
+ GRPO w/ GRRM	67.38	88.42	64.15	85.19	63.40	78.88	68.57	85.38	65.92	79.35
GRRM with Instruct-Tuned Base Model										
Qwen2.5-7B-Instruct	67.31	85.84	59.92	79.42	58.72	70.84	66.59	79.95	62.75	73.93
+ GRPO w/ GRRM	67.15	89.40	64.88	87.07	63.89	82.44	68.74	86.99	67.37	83.45

Table 8: **Additional ablation results.** We report performance for (1) Applying different reward models to a non reasoning model tuned from `Qwen2.5-7B`, and (2) Applying GRRM optimization on top of the instruction-tuned `Qwen2.5-7B-Instruct` model.

of grass), the model identifies the underlying semantic meaning of dissatisfaction and envy. It then retrieves the culturally equivalent Chinese proverb "这山望着那山高" (*This mountain looks higher than that one*), ensuring the translation resonates with native speakers.

Similarly, in the **Decoding Internet Slang** (Case 2), the model encounters the phrase "INTJ总是装E", which blends technical MBTI terminology with colloquial Chinese slang. A literal breakdown might fail to capture the nuance of "装" (feign/pretend) in this specific context. However, the model's reasoning trace explicitly decomposes the acronyms and analyzes the character traits, correctly deriving the idiomatic English translation: "putting on an extroverted front."

These cases confirm that our approach does not merely memorize translation pairs but actively performs semantic analysis and cultural alignment, justifying the substantial improvements seen in the LLM-judge evaluations.

F GRRM Application: Inference-time Reranking

In this section, we investigate the effectiveness of GRRM as a verifier to select the best translation from multiple candidates during inference.

Experimental Setup. We conduct experiments on both reasoning and non-reasoning models, evaluating their respective SFT baselines and GRPO-optimized checkpoints. For candidate generation, we employ standard sampling with a temperature of 0.6 to produce $N = 4$ candidates per input. We compare our GRRM reranking strategy against: (1) **Sampling Baseline**, representing the expected performance of the policy; and (2) **Best-of-N**, which selects the candidate with the highest log-probability assigned by the policy model. Additionally, for non-reasoning models, we include **Beam Search** with a beam width of 4 as a strong decoding baseline. Notably, we exclude Beam Search for reasoning models, as the extensive length of

Base Model	Inference Strategy	WMT23		WMT24++		Seed-X-Challenge				Average	
		Zh→En		En→Zh		Zh→En		En→Zh			
		BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge	BLEURT	oss-judge
SFT Baseline (w/ Reasoning)	Sampling Baseline	66.78	85.04	59.33	74.20	66.95	80.17	61.64	70.02	63.67	77.36
	Best-of-N	66.76	85.21	59.79	75.50	66.88	79.57	61.94	71.27	63.84	77.89
	Ranking w/ GRRM	**67.68**	**88.06**	**61.50**	**81.47**	**68.05**	**83.22**	**63.29**	**77.15**	**65.13**	**82.47**
+ GRPO w/ GRRM	Sampling Baseline	67.40	90.15	64.77	88.05	69.23	87.41	67.20	84.79	67.15	87.60
	Best-of-N	**67.55**	90.28	64.99	88.26	**69.42**	87.73	67.21	83.89	67.29	87.54
	Ranking w/ GRRM	67.45	**90.41**	**65.17**	**88.90**	69.16	**87.77**	**67.54**	**85.90**	**67.33**	**88.25**
SFT Baseline (w/o Reasoning)	Sampling Baseline	66.05	82.63	61.50	74.63	64.16	73.75	63.04	69.29	63.69	75.07
	Best-of-N	67.11	84.49	62.42	77.48	64.88	73.85	62.91	71.00	64.33	76.71
	Beam Search	**68.18**	86.28	**63.62**	**81.34**	**66.25**	77.68	**64.16**	75.45	**65.55**	80.19
	Ranking w/ GRRM	67.54	**86.65**	63.18	81.06	65.75	**78.35**	63.99	**75.46**	65.12	**80.38**
+ GRPO w/ GRRM	Sampling Baseline	67.17	88.10	63.91	84.93	68.19	84.97	65.96	79.44	66.31	84.36
	Best-of-N	67.38	88.47	64.12	85.11	68.47	85.24	65.66	78.00	66.40	84.21
	Beam Search	**67.54**	88.81	64.20	85.45	68.48	85.62	65.98	78.64	66.55	84.63
	Ranking w/ GRRM	67.37	**89.21**	**64.32**	**86.27**	**68.53**	**86.59**	**66.60**	**80.03**	**66.70**	**85.53**

Table 9: **Inference-time Reranking Performance.** We compare different decoding strategies across reasoning and non-reasoning models.

Chain-of-Thought sequences renders the decoding process prohibitively slow and computationally intractable.

Results and Analysis. The results presented in Table 9 show that ranking with GRRM consistently outperforms the Sampling and Best-of-N baselines across all datasets and model stages. Even for GRPO-optimized models, which have already been aligned via RL, inference-time reranking provides further performance gains. For reasoning models where Beam Search is computationally intractable, *scaling test-time compute* via GRRM reranking proves to be a robust inference paradigm. Rather than relying on a single generation, this verify-and-select strategy significantly boosts the SFT baseline (e.g., an average increase of over 5 points in LLM-judge scores) by effectively filtering out flawed reasoning paths. In non-reasoning settings, while Beam Search remains competitive on BLEURT, GRRM reranking consistently achieves superior LLM-judge scores. This suggests that GRRM prioritizes semantic fidelity and better aligns with human preferences compared to likelihood-based decoding strategies.

GRRM Prompt Example

User: Given a source text in English and multiple translation candidates in Chinese. Perform a step by step analysis and comparison of the translation quality for the candidates. Finally, rank and score the candidates with integer scores on a scale from 0 to 10.

Source text:
“‘
Garland is the chief judge for the U.S. Court of Appeals for the District of Columbia Circuit.
“‘

Translation A:
“‘
加兰是美国哥伦比亚特区上诉法院首席法官。
“‘
Translation B:
“‘
加兰德是美国哥伦比亚特区联邦巡回上诉法院的首席法官。
“‘

GRRM Output

Analysis and Comparison

Source Text: Garland is the chief judge for the U.S. Court of Appeals for the District of Columbia Circuit.

1. **Translation A:** 加兰是美国哥伦比亚特区上诉法院首席法官。
    * **Accuracy:** This translation correctly identifies Garland as the chief judge of a court in the D.C. area, but it is imprecise. The "Court of Appeals for the District of Columbia Circuit" is a specific federal appellate court in the U.S. legal system. "District of Columbia Circuit" is a key part of its name that signifies it’s part of a "Circuit" of federal appeals courts, not just any court in D.C. By translating "Circuit" as "上诉法院" (Court of Appeals), the translation implies he is a chief judge of the D.C. "appellate court" generally, which misses the "Circuit" designation that distinguishes it from other appellate courts. This is a loss of specific legal precision.
    * **Clarity:** The meaning is generally clear.

2. **Translation B:** 加兰德是美国哥伦比亚特区联邦巡回上诉法院的首席法官。
    * **Accuracy:** This translation is highly accurate. "联邦巡回上诉法院" (Federal Circuit Court of Appeals) is the standard official name for "U.S. Court of Appeals" in the context of U.S. federal courts. The structure "Circuit Court of Appeals for the District of Columbia" implies it’s the D.C. court of this specific federal judicial "Circuit". This translation correctly identifies Garland as the chief judge of the full, specific court with the correct legal terminology in Chinese.
    * **Clarity:** The meaning is unambiguous and fully preserves the nuances of the original text. "加兰德" is the standard pinyin.

Conclusion

Translation A is acceptable but lacks specificity due to the omission of "Circuit" in the court’s name. It’s understandable but not technically correct in a legal context.

Translation B is excellent. It uses the precise and official name "联邦巡回上诉法院" for the "U.S. Court of Appeals", correctly translating "Circuit" as part of the full name of the court. It is both accurate and clear.

Final Ranking:

B > A

Scores:

B: 10, A: 7

Figure 6: An example of GRRM for machine translation evaluation. Given a source sentence and multiple translation candidates, the model performs comparative reasoning, then outputs a ranking and relative integer scores. The GQM prompt used for data annotation is similar, except for the additional requirements regarding the output format.

Case 1: Translating Idiomatic Expressions (En → Zh)

Source: The grass is always greener on the other side.

Model Reasoning Trace

Step-by-step Analysis

1. **"The grass is always greener on the other side."**: This is a well-known English idiom. A literal, word-for-word translation would be unnatural in Chinese. The core meaning is that people often feel their current situation is not as good as they perceive another person's or another situation to be.

2. **Identifying the Chinese equivalent**: The most common and idiomatic translation for this concept in Chinese is "这山望着那山高" (zhè shān wàngzhe nà shān gāo), which literally means "To look at the other mountain, thinking it's higher." This perfectly captures the essence of wanting what one doesn't have.

3. **Exploring other options**:
* "彼岸的草总是更绿" (Bǐ'àn de cǎo zǒngshì gèng lǜ): A more literal, poetic translation, "The grass on the other shore is always greener," which is also understandable but not as common in everyday speech as the idiom.
* "隔岸观火" (gé àn guān huǒ): This idiom means "to watch a fire from the other bank," implying a passive, detached, or envious observation, but it doesn't fully capture the comparison between one's own situation and another's.

4. **Final Selection**: "这山望着那山高" is the most idiomatic and widely used equivalent. A slightly more formal and written version is "人往高处走，水往低处流" (rén wǎng gāochù zǒu, shuǐ wǎng dīchù liú), meaning "People strive for higher ground, water flows to lower ground," which also conveys a similar message. However, "这山望着那山高" is the most direct and natural translation.

Final Translation

这山望着那山高。

Case 2: Decoding Internet Slang (Zh → En)

Source: INTJ总是装E

Model Reasoning Trace

Step-by-Step Analysis

1. **"INTJ"**: This is an acronym for a specific personality type in the Myers-Briggs Type Indicator (MBTI), which stands for "Introverted, Intuitive, Thinking, Judging". It is a technical term and should be kept as "INTJ".

2. **"总是"**: This means "always", "constantly", or "tend to always". It implies a consistent, habitual behavior.

3. **"装E"**: This is the core of the phrase.
* "装" (zhuāng) means "to pretend", "to act", "to feign", "to put on an act", or "to be fake".
* "E" is an abbreviation for "Extroverted" (from the Myers-Briggs Type Indicator's "Extraverted" trait).
* Combined, "装E" means "to act like an extrovert" or "to put on an extroverted front". The person is pretending to be outgoing and sociable when they are, in reality, more introverted.

4. **Synthesis**: A literal translation would be "INTJ always pretends to be an extrovert". A slightly more natural and idiomatic version would be "INTJs are always putting on an extroverted front" or "INTJs constantly feign extroversion". "INTJs always act like extroverts" is also a good option.

5. **Final Selection**: "INTJs are always putting on an extroverted front" is a very accurate and natural translation that captures the essence of the original phrase.

Final Translation

INTJs are always putting on an extroverted front.

Figure 7: Qualitative analysis of translation reasoning. We showcase the model's ability to handle cultural nuances by identifying **idiomatic equivalents** (Case 1) and **decoding internet slang** (Case 2) through step-by-step analysis.