Diversity-Aware Literary Machine Translation with Multi-Reward Policy Optimization

Zeynep Yirmibeşoğlu Balal **Tunga Güngör**
Computer Engineering, Boğaziçi University, Istanbul, Türkiye
zeynep.yirmibesoglu@std.bogazici.edu.tr
gungort@bogazici.edu.tr

Abstract

Literary translation is a difficult task that not only requires semantic accuracy but also stylistic richness and lexical diversity. Pretrained and supervised fine-tuned Large Language Models (LLMs) can over-rely on safe vocabulary choices, leading to translations that lack lexical variety. To address this problem, we propose a novel diversity-aware multi-objective Group Relative Policy Optimization (GRPO) framework that pushes the limits of open-source translation quality while increasing lexical diversity. We introduce two diversity-aware reward mechanisms, a Leave-One-Out (LOO) marginal contribution reward and a Self-BLEU penalty, balanced alongside neural quality metrics (COMET), lexical overlap (BLEU), and structural constraints. Through experiments on Turkish–English and German–English using Qwen3-14B, we show that our diversity-aware reinforcement learning approach successfully enhances lexical richness alongside translation quality. Our models achieve state-of-the-art open-source performance in literary translation, bridging the gap with leading commercial systems and demonstrating that policy optimization can effectively steer LLMs toward high-quality, lexically diverse outputs.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in Machine Translation (MT), but literary translation still remains a demanding challenge (Wu et al., 2025; Scalena et al., 2026). Unlike technical or news translation, literary texts require the preservation of stylistic nuances, cultural context, and a rich, varied vocabulary (Voigt and Jurafsky, 2012). A successful literary translation must not only be semantically faithful to the source but also exhibit the linguistic creativity inherent in human writing (Guerberof-Arenas and Toral, 2022). In this research, we focus on lexical diversity and translation quality of open-source LLMs in the literary translation task.

The standard paradigm for adapting LLMs to translation tasks relies heavily on Supervised Fine-Tuning (SFT). While SFT is highly effective at teaching models to follow translation instructions, models can sometimes default to linguistic simplification, prioritizing high-probability, safe token sequences. As a result, the fine-tuned models may produce translations that lack the full lexical diversity of the original texts.

Reinforcement Learning (RL) techniques, particularly critic-free methods like Group Relative Policy Optimization (GRPO) (Shao et al., 2024), offer a promising alternative by allowing models to optimize directly for sequence-level objectives. In this paper, we introduce a diversity-aware GRPO framework suitable for the creativity demands of literary MT. We propose a multi-objective reward structure that balances semantic fidelity and lexical accuracy with explicit emphasis on linguistic variety. Specifically, we integrate two novel diversity rewards: a Leave-One-Out (LOO) reward that quantifies a generation's unique marginal contribu-

tion to a sampled group, and a Self-BLEU penalty that discourages n-gram overlap among candidate translations. These are combined with COMET, BLEU, and a length penalty to ensure the model remains anchored to high-quality, structurally sound outputs.

We evaluate our framework across two language pairs (Turkish–English and German–English) in both directions using Qwen3-14B. Our comprehensive analysis, incorporating both neural quality metrics and lexical richness measures (MTLD, Yule's I, TTR), reveals that our diversity-aware GRPO approach not only pushes the boundaries of open-source model performance but also significantly enhances lexical richness. By explicitly rewarding diversity, our models achieve high translation quality, reduce critical translation errors, and rival the quality and lexical variance of proprietary models like GPT-5.2. The code and the dataset splits are available on Github[1] except for the Turkish–English dataset, which is not publicly available due to copyright issues.

2 Related Work

2.1 Evolution of Policy Optimization

The paradigm for adapting LLMs has shifted significantly from SFT toward RL techniques designed to capture complex human preferences and reasoning constraints. One of the early efforts is Proximal Policy Optimization (PPO) (Schulman et al., 2017), which established the Reinforcement Learning from Human Feedback (RLHF) standard using a clipped surrogate objective for stability; however, its four-model architecture poses significant scaling bottlenecks. Improving efficiency, Direct Preference Optimization (DPO) (Rafailov et al., 2024) bypassed explicit reward models by optimizing the policy directly using preference data. Subsequent refinements include KTO (Ethayarajh et al., 2024), which uses unpaired binary feedback, and SimPO (Meng et al., 2024), which introduces reference-free objectives to mitigate length-bias.

The need for models to learn from their own actively generated data has driven the rise of critic-free online methods. GRPO (Shao et al., 2024) eliminates the value function (critic), instead sampling a group of outputs per input and using the group's average reward as a baseline. This balances online feedback and memory efficiency, which became a foundation for state-of-the-art reasoning models. Refinements like REINFORCE++ (Hu et al., 2025) propose Global Advantage Normalization to reduce bias in GRPO's local group normalization, improving generalization.

2.2 Policy Optimization in Machine Translation

In MT, policy optimization has evolved toward enforcing complex stylistic constraints, prioritizing sequence-level metrics over standard cross-entropy. Early efforts like Contrastive Preference Optimization (CPO) (Xu et al., 2024) addressed the "adequacy-fluency" trade-off by training models to distinguish between high-quality references and imperfect outputs. Direct Quality Optimization (DQO) (Uhlig et al., 2025) utilized CometKiwi22 as a proxy to align model outputs with human preferences.

GRPO has catalyzed "reasoning-centric" translation involving internal planning and self-correction. Frameworks like R1-T1 (He et al., 2025) and MT-R1-Zero (Feng et al., 2025) leverage GRPO to encourage hierarchical translation strategies. SSR-Zero (Yang et al., 2025b) extends this by letting the model act as its own judge, removing dependency on external references.

GRPO is also effective for enforcing lexical constraints. Garcia Gilabert et al. (2025) adapted models using monolingual data by optimizing a joint reward for terminology and quality. Similarly, TAT-R1 (Li et al., 2025) integrated word alignment rewards to incentivize accurate terminology. These methods focus on token-level accuracy rather than broader sequence diversity.

A persistent challenge in GRPO is "reward hacking", where optimizing metrics like BLEU leads to safe, repetitive patterns sacrificing lexical richness. To mitigate this, MO-GRPO (Ichihara et al., 2025) and AW-GRPO (Ichihara and Jinnai, 2025) propose dynamic reward normalization and weight adjustment for stable learning. While improving stability, maintaining lexical diversity in specialized domains like literary translation remains largely under-explored.

2.3 Diversity in Policy Optimization

Despite the success of GRPO, models often suffer from "mode collapse", where the policy converges on a narrow subset of high-probability responses. To mitigate this, Group-Aware Policy

[1] https://github.com/zeynepyirmibes/mt_multi_objective_optimization_unsloth

Optimization (GAPO) (Anschel et al., 2025) rewards sampled groups as a whole to encourage uniform coverage. Similarly, DRA-GRPO (Chen et al., 2026) addresses "diversity-quality inconsistency" by downweighting redundant completions, while MMR-GRPO (Wei and Huang, 2026) leverages Maximal Marginal Relevance to prioritize variety and accelerate convergence. SetPO (Li et al., 2026) introduces a leave-one-out marginal contribution mechanism, rewarding each trajectory based on its unique impact on the set's diversity. While these efforts improve reasoning in objective domains like mathematical reasoning and code generation, the use of policy optimization to intentionally broaden the lexical variety of translations remains a significant and under-explored challenge.

3 Methodology

3.1 Group Relative Policy Optimization (GRPO)

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) enhances LLM reasoning while reducing resource overhead by eliminating the value function (critic). For each query q, GRPO samples a group of G outputs $\{o_1, \ldots, o_G\}$ from the current policy $\pi_{\theta_{\text{old}}}$. It uses the group average reward as the baseline by utilizing the group's reward statistics to estimate the advantage $\hat{A}_i$ for the i-th output:

$$\hat{A}_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G) + \epsilon} \tag{1}$$

The policy π_θ is then updated by maximizing a clipped surrogate objective, regularized by a Kullback-Leibler (KL) divergence term against a reference model π_{ref}:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Big(\mathcal{L}_i^{\text{clip}}(\theta) \\ - \beta \mathbb{D}_{KL}(\pi_\theta \| \pi_{\text{ref}}) \Big) \Bigg] \\ \mathcal{L}_i^{\text{clip}}(\theta) = \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \Big(\rho_{i,t} \hat{A}_i, \\ \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon) \hat{A}_i \Big) \end{aligned} \tag{2}$$

where $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$ represents the probability ratio. The coefficient β controls the strength of the KL penalty, ensuring the optimized policy remains within a stable trust region of the reference distribution.

3.2 Diversity-Aware Reward Modeling

To effectively steer the model toward high-quality and lexically diverse outputs, we design a multi-objective reward structure. While standard RL training can safely rely on automated translation quality metrics, this can lead to "mode collapse" where the model prefers safe, repetitive phrasing. To mitigate this, we propose two novel diversity reward functions designed to reward lexical and structural variety within the sampled group G. These diversity rewards are coupled with established translation quality functions (COMET and BLEU) and a length penalty to prevent the model from hacking the diversity reward by generating overly verbose outputs. By balancing these objectives, the reward signal encourages the model to explore the semantic space more broadly, ensuring that the resulting translations are not only accurate but also lexically creative and diverse, reflecting the natural variation found in human translation.

Leave-One-Out (LOO) Diversity Reward To encourage lexical variety, we propose a Leave-One-Out (LOO) Diversity Reward. Unlike standard diversity metrics that provide a single score for an entire group, the LOO reward quantifies the marginal contribution of an individual completion o_i to the collective diversity of its group G. We define the diversity of a set of completions using the Distinct-2 score (Li et al., 2016), which is the ratio of unique bigrams to total bigrams. Let $B(o_i)$ be the set of bigrams in completion o_i, and $B(G) = \bigcup_{j=1}^{|G|} B(o_j)$ be the multiset of all bigrams in the group. The group diversity $D(G)$ is defined as $D(G) = \frac{|\text{unique}(B(G))|}{|B(G)|}$.

The marginal contribution M_i of completion o_i is then calculated by comparing the diversity of the full group to the diversity of the group excluding o_i: $M_i = D(G) - D(G \setminus \{o_i\})$.

Finally, the reward $R_{\text{LOO}}(o_i)$ is centered at a neutral value of 0.5 and scaled to provide a clear gradient signal:

$$R_{\text{LOO}}(o_i) = \text{clip}\left(0.5 + \alpha \cdot M_i, 0, 1\right) \tag{3}$$

where α is a scaling factor (e.g., $\alpha = 5$) that amplifies the small marginal differences. This formulation ensures that completions introducing unique

vocabulary relative to their peers receive a reward > 0.5, while redundant completions that merely repeat n-grams already present in the group are penalized with a reward < 0.5.

Self-BLEU Diversity Reward While the LOO reward captures the marginal contribution to the group's vocabulary, we also implement a Self-BLEU diversity reward to explicitly penalize n-gram overlap between individual completions. Self-BLEU (Zhu et al., 2018) is an established metric for evaluating the diversity of generated text, serving as a measure of how much a specific output mirrors the rest of the generated samples.

For each completion o_i within a group G, we calculate its similarity to all other completions o_j ($j \neq i$) by treating o_i as the hypothesis and o_j as the reference. The Self-BLEU score for o_i is defined as the average sentence-level BLEU score across all such pairs:

$$\text{Self-BLEU}(o_i) = \frac{1}{|G| - 1} \sum_{j \neq i} \text{BLEU}(o_i, o_j) \tag{4}$$

To transform this similarity metric into a maximization objective for GRPO, we define the reward $R_{\text{Self-BLEU}}(o_i)$ as the complement of the similarity score: $R_{\text{Self-BLEU}}(o_i) = 1 - \text{Self-BLEU}(o_i)$.

This formulation yields a reward of 1.0 for a completion that is entirely unique within its group and approaches 0.0 for completions that are nearly identical to their peers. By integrating this reward, we provide the model with a clear signal to avoid redundant, safe translations and instead explore diverse linguistic structures.

Translation Quality Rewards While diversity rewards encourage linguistic exploration, the primary objective remains the generation of accurate and fluent translations. To ensure that the model does not sacrifice fidelity for variety, we incorporate two complementary metrics as quality rewards: BLEU for lexical overlap and COMET for semantic similarity.

We utilize the *sacrebleu* implementation to compute sentence-level BLEU scores. To maintain consistency with the [0, 1] reward space required for stable RL training, we normalize the score by dividing the BLEU score by 100.

To capture nuances that n-gram metrics often miss, we also use COMET[2]. Since COMET scores can occasionally fall outside the standard range for extremely poor or high-quality translations, we apply a clamping function to ensure numerical stability during the GRPO update and clip COMET scores between 0 and 1. By combining these two metrics, the reward signal provides a balanced view of quality, rewarding both exact word matches and high-level semantic equivalence.

Log-Ratio Length Penalty A common challenge in optimizing for diversity via RL is reward hacking, where the model learns to exploit the reward structure without improving task performance. In our early experiments, we observed that the model attempted to maximize the diversity rewards by significantly increasing the length of the generated translations, as longer sequences naturally contain more unique n-grams. This not only degraded translation quality but also led to substantial increases in training time and computational overhead.

To mitigate this, we introduce a Log-Ratio Length Penalty (R_{len}). This penalty is designed to be symmetric, penalizing translations that are significantly longer or shorter than the reference translation with equal weight. For a generated translation o_i with word count l_{gen} and a reference translation y with word count l_{ref}, the penalty is calculated as:

$$R_{\text{len}}(o_i) = \max\left(-\left|\ln\left(\frac{l_{\text{gen}}}{l_{\text{ref}}}\right)\right|, -2.0\right) \tag{5}$$

This "V-shaped" penalty function offers several advantages: 1) A translation that is twice as long as the reference is penalized exactly as much as the one that is half as long (approximately -0.7); 2) Small, natural variations in length (e.g., $\pm 10\%$) incur negligible penalties, allowing the model to retain linguistic flexibility; 3) By capping the penalty at -2.0, we prevent extreme outliers from overdominating the gradient signal and drowning out the quality rewards from BLEU or COMET.

Reward Aggregation and Weighting The final reward R_{total} is a weighted sum of the individual components described above. We consolidate our objectives into a single signal to ensure that the policy simultaneously optimizes for lexical accuracy, semantic fidelity, and diversity, while remaining constrained by length requirements. Based on empirical tuning during our initial experiments, we

[2]Unbabel/wmt22-comet-da

assigned a specific weight to each reward component to prioritize translation quality while maintaining a sufficient gradient signal for diversity. The total reward is defined as:

$$R_{\text{total}} = w_1 R_{\text{BLEU}} + w_2 R_{\text{COMET}} + w_3 R_{\text{div}} + w_4 R_{\text{len}} \quad (6)$$

where R_{div} represents either the LOO-Diversity or the Self-BLEU reward, depending on the experimental configuration. Reward weights used in our experiments are empirically selected, balancing quality and diversity for all language directions (Table 1).

Reward	Weight (w)	Objective
R_{BLEU}	0.3	Lexical Overlap
R_{COMET}	0.4	Semantic Fidelity
R_{div}	0.2	Lexical Diversity
R_{len}	0.1	Structural Constraint

Table 1: Empirically determined weights for the composite GRPO reward function.

By assigning the highest weight to COMET (0.4), we prioritize the semantic integrity of the translation, as neural-based metrics have been shown to correlate more highly with human judgment than n-gram metrics alone (Rei et al., 2020). The diversity component is weighted at 0.2, providing enough influence to prevent mode collapse without allowing the model to prioritize uniqueness over actual translation accuracy. Finally, the length penalty acts as a regularization term (0.1) to ensure the outputs remain concise and structurally aligned with human references.

4 Experimental Setup

4.1 Datasets

Literary translation poses unique challenges for machine translation due to its rich vocabulary, complex syntactic structures, and the prevalence of figurative language. While literal, word-for-word translation is often acceptable or even preferred in news and technical domains to ensure factual precision, literary texts require the preservation of stylistic nuances, metaphors, and cultural context. This is why we choose the literary domain for evaluating the ability of policy-optimized MT models to balance translation fluency with lexical diversity.

We conduct experiments on two language pairs in both directions: Turkish–English (**TR-EN**) and German–English (**DE-EN**). For **TR-EN**, we utilize a manually curated corpus of 93 aligned literary works (Yirmibeşoğlu et al., 2023). For **DE-EN**, we use the *Books* corpus from the OPUS collection (Tiedemann, 2012).

To keep the integrity of our evaluation, we split the data at the book level, ensuring that no individual book appears in both the training and test sets. We use 50,000 (TR-EN)/30,000 (DE-EN) sentences for SFT and 5,000 for GRPO (see Table 7 in the Appendix for full statistics). The 5,000 sentences we use for GRPO training are entirely disjoint from the SFT training set. We perform deduplication and normalize punctuation using a custom dictionary. We use the native tokenizer of our base model, Qwen3-14B.

4.2 Models

To evaluate the effectiveness of policy optimization in the literary domain, we establish a baseline for open-source performance by performing zero-shot inference on LLaMA-3.1-8B (Grattafiori et al., 2024), Qwen3-8B, and Qwen3-14B (Yang et al., 2025a). Among these, Qwen3-14B demonstrated superior translation quality in our preliminary evaluations for both language pairs. Thus, we select Qwen3-14B as our base model for the subsequent SFT and GRPO experiments. For all evaluations and model generations, we employ greedy decoding to ensure deterministic and reproducible outputs.

For a performance ceiling, we evaluate the zero-shot performance of proprietary frontier models, specifically GPT-4o and GPT-5.2 with no reasoning (API versions as of February 2026). All baselines, both open-source and proprietary, are evaluated using the identical system prompt detailed in Appendix B to ensure experimental control.

We initialize our experiments with Qwen3-14B without reasoning, using SFT as a "cold-start" phase to adapt the model to specific language directions and output format instructions. GRPO is subsequently performed on top of the SFT-trained model. Both phases utilize the Unsloth library for memory efficiency, employing Rank-Stabilized LoRA (RS-LoRA) (Kalajdzievski, 2023) across all linear layers. For the GRPO phase, we initialize the model with the adapter weights from the previous SFT stage. We set the KL divergence param-

eter $\beta = 0$; this allows the policy to deviate from the reference model to intentionally maximize the model's capacity for lexical diversity during optimization. Detailed hyperparameters are provided in Appendix C.

We use the reward weights specified in Table 1 for our primary GRPO experiments. Additionally, we perform a comparison run using only BLEU (0.4), COMET (0.5), and the length penalty (0.1), where the weight previously allocated to the diversity reward is distributed evenly between the BLEU and COMET components.

4.3 Evaluation

To provide a comprehensive assessment of translation performance, we utilize both standard quality metrics and lexical diversity measures. We evaluate translation quality using BLEU (via *sacreBLEU*), BLEURT (*BLEURT-20*), and COMET. For reference-free evaluation of quality, we use COMET-Kiwi (*wmt22-cometkiwi-da*). We also conduct error analysis with XCOMET (*XCOMET-XXL*). We normalize the total number of translation errors for each error category (minor, major, and critical) with the number of sentences in the test set and report the error rates, considering only annotations with $> 50\%$ confidence.

Given that lexical diversity is a crucial indicator of how well a model captures the richness of literary vocabulary and avoids repetitive outputs, we further analyze the results using three distinct metrics computed via the LexicalRichness module (Shen, 2022): Type-Token Ratio (TTR), Yule's I, and the Measure of Textual Lexical Diversity (MTLD). TTR provides a fundamental ratio of unique types to total tokens (Templin, 1957), while Yule's I offers a more stable probability-based model of word frequency distributions that is less sensitive to varying text lengths (Yule, 1944). Finally, we include MTLD, which calculates the average length of text sequences that maintain a specific TTR threshold, providing the most robust measure to text length (McCarthy, 2005).

5 Results and Analysis

We present our main experimental results and provide ablation studies.

5.1 Main Results

We present a comprehensive evaluation of our proposed GRPO framework across four translation directions: Turkish–English (TR-EN), English–Turkish (EN-TR), German–English (DE-EN), and English–German (EN-DE). We compare our models against state-of-the-art proprietary baselines (GPT-4o, GPT-5.2) and open-source models (LLaMa-3.1, Qwen3). Tables 2 through 5 summarize the results using standard translation quality metrics (BLEU, COMET, COMET-Kiwi, BLEURT) and lexical diversity metrics (TTR, Yule's I, MTLD). In each table the results are grouped according to training setting: pretrained models as baselines, supervised fine-tuned Qwen3-14B, and GRPO-trained Qwen3-14B models.

Baseline Models A consistent trend across all four directions is the superiority of proprietary models, particularly GPT-5.2, in terms of both translation quality and lexical richness. Dominance of GPT-5.2 in lexical diversity (MTLD) is common across all language directions, while translation quality supremacy is shared between GPT-4o and GPT-5.2 within the pretrained group. For EN-TR (Table 3), GPT-4o (1027.3) and GPT-5.2 (1168.0) achieve a remarkably high MTLD score, significantly outperforming all zero-shot open-source models. This suggests these large-scale proprietary models maintain a more varied vocabulary during the translation process, especially for the morphologically-rich Turkish target language, whereas smaller open-source models tend toward more conservative, safe linguistic choices in zero-shot setting. Among open-source baselines, Qwen3-14B consistently emerges as the strongest performer in translation quality, serving as a competitive foundation for further fine-tuning.

Lexical Diversity in SFT Our results highlight a significant trade-off in standard SFT. While the Qwen3-14B (SFT) model shows marked improvements in BLEU scores, it suffers a substantial decline in lexical diversity for all language directions except EN-TR. MTLD score drops from 149.2 to 96.3 for TR-EN, from 120.2 to 91.6 for DE-EN, and from 203.9 to 142.5 for EN-DE. In the EN-TR direction, the MTLD is already too low (62.7) in the base model, which rises to 166.8 in the SFT-trained counterpart. While SFT is crucial for teaching models to follow instructions and match reference styles, it does not inherently optimize for linguistic richness and can lead to an over-reliance on high-probability tokens. This motivates our use

	Model	BLEU	COMET	Kiwi	BLEURT	TTR	Yule's I	MTLD
Pretrained	GPT-4o	**17.82**	0.792	0.827	0.641	0.220	6.36	150.2
	GPT-5.2	16.34	**0.795**	**0.830**	**0.643**	0.226	6.95	**162.9**
	LLaMa-3.1	13.55	0.761	0.801	0.595	**0.231**	**7.10**	147.0
	Qwen3-8B	13.90	0.765	0.807	0.603	0.203	4.61	41.5
	Qwen3-14B	15.53	0.783	0.819	0.622	0.212	5.86	149.2
SFT	Qwen3-14B (SFT)	18.72	0.771	0.786	0.607	0.202	4.05	96.3
GRPO	Qwen3-14B (GRPO: B+C+LP)	**21.23**	0.805	0.819	0.645	0.217	5.25	133.1
	Qwen3-14B (GRPO: B+C+L+LP)	19.44	0.802	0.817	0.642	**0.234**	**6.87**	**154.6**
	Qwen3-14B (GRPO: B+C+S+LP)	21.10	**0.806**	**0.820**	**0.648**	0.226	6.29	145.2

Table 2: Model Performance Comparison for TR-EN Translation. GRPO rewards are abbreviated as B (BLEU), C (COMET), S (Self-BLEU), L (Leave-One-Out / LOO diversity), and LP (Length Penalty). Kiwi refers to COMET-Kiwi. Highest scores in the first and third groups are shown in **bold**.

	Model	BLEU	COMET	Kiwi	BLEURT	TTR	Yule's I	MTLD
Pretrained	GPT-4o	**9.51**	0.786	0.821	0.629	0.468	59.53	1027.3
	GPT-5.2	9.04	**0.795**	**0.827**	**0.637**	**0.485**	**71.71**	**1168.0**
	LLaMa-3.1	5.97	0.735	0.756	0.553	0.454	48.75	439.4
	Qwen3-8B	5.37	0.734	0.761	0.554	0.388	22.72	36.4
	Qwen3-14B	7.17	0.757	0.788	0.584	0.418	38.42	62.7
SFT	Qwen3-14B (SFT)	9.50	0.763	0.766	0.586	0.431	60.66	166.8
GRPO	Qwen3-14B (GRPO: B+C+LP)	**9.96**	0.782	0.799	0.610	0.434	56.26	201.8
	Qwen3-14B (GRPO: B+C+L+LP)	7.23	0.799	0.793	0.622	**0.478**	**100.38**	**1101.9**
	Qwen3-14B (GRPO: B+C+S+LP)	9.04	**0.805**	**0.807**	**0.632**	0.457	74.25	1086.3

Table 3: Model Performance Comparison for EN-TR Translation

of further optimization on the fine-tuned Qwen3-14B model: to explicitly push the limits of both translation quality and lexical diversity.

GRPO: Restoring Quality and Diversity GRPO training on top of the SFT model yields immediate benefits. Even the baseline GRPO configuration (B+C+LP) successfully enhances lexical diversity compared to the SFT baseline. For TR-EN, EN-TR, and DE-EN, this configuration achieves the highest BLEU scores, surpassing the commercial GPT models. Crucially, even without explicit diversity rewards (leave-one-out and Self-BLEU), the GRPO process improves MTLD compared to the SFT baseline (e.g., from 96.3 to 133.1 in TR-EN). This suggests that the relative rewarding mechanism of GRPO helps the model explore a more effective output space than the rigid teacher-forcing of SFT.

The Impact of Diversity Rewards The introduction of explicit diversity rewards—Self-BLEU (S) and Leave-One-Out (L)—further pushes the boundaries of model performance. In EN-TR, the B+C+L+LP configuration with the Leave-One-Out diversity reward achieves an MTLD of 1101.9, nearly matching the GPT-5.2 baseline and drastically outperforming the SFT model (166.8). Across all language directions except EN-DE, the GRPO-optimized Qwen3-14B—leveraging diversity rewards—attained the highest MTLD scores among open-source configurations. In most cases, adding a diversity reward does not degrade translation quality; instead it often improves it. In TR-EN and DE-EN, the B+C+S+LP setup with the Self-BLEU reward reaches the highest BLEURT and COMET scores, surpassing the GPT models, while achieving higher or comparable lexical richness.

Overall, the results demonstrate that while SFT is necessary for instruction following, it does not inherently optimize for the creativity and lexical range of the translation. Our proposed method of GRPO with multi-objective reward function—specifically one that rewards linguistic variety—provides a robust solution that balances translation quality and lexical diversity. We observe that GRPO: B+C+S+LP generally offers the best balance, delivering state-of-the-art open-source literary translation quality while simultaneously closing the diversity gap between open-source models and proprietary models.

Error Analysis We conduct error analysis on the pretrained Qwen3-14B, and its SFT- and GRPO-trained counterparts using XCOMET (Figure 1).

	Model	BLEU	COMET	Kiwi	BLEURT	TTR	Yule's I	MTLD
Pretrained	GPT-4o	**22.19**	**0.787**	**0.805**	**0.643**	0.187	3.91	121.7
	GPT-5.2	20.97	0.776	**0.805**	0.642	**0.195**	**4.45**	**126.8**
	LLaMa-3.1	19.42	0.766	0.795	0.626	0.187	3.93	119.3
	Qwen3-8B	19.85	0.770	0.798	0.629	0.184	3.83	121.7
	Qwen3-14B	21.11	0.777	0.802	0.637	0.181	3.66	120.2
SFT	Qwen3-14B (SFT)	25.72	0.760	0.756	0.624	0.175	3.09	91.6
GRPO	Qwen3-14B (GRPO: B+C+LP)	**28.20**	**0.789**	0.788	**0.656**	0.187	3.68	111.4
	Qwen3-14B (GRPO: B+C+L+LP)	26.91	0.785	0.790	0.650	0.191	3.98	122.0
	Qwen3-14B (GRPO: B+C+S+LP)	27.62	**0.789**	**0.791**	**0.656**	**0.194**	**4.27**	**127.1**

Table 4: Model Performance Comparison for DE-EN Translation

	Model	BLEU	COMET	Kiwi	BLEURT	TTR	Yule's I	MTLD
Pretrained	GPT-4o	16.72	0.766	0.814	0.639	0.242	10.11	198.1
	GPT-5.2	**16.96**	**0.773**	**0.815**	**0.658**	**0.251**	**11.40**	**209.5**
	LLaMa-3.1	14.84	0.737	0.791	0.610	0.239	9.47	179.1
	Qwen3-8B	12.83	0.744	0.800	0.616	0.244	10.66	203.6
	Qwen3-14B	14.21	0.753	0.809	0.631	0.247	10.86	203.9
SFT	Qwen3-14B (SFT)	18.44	0.739	0.750	0.625	0.227	7.53	142.5
GRPO	Qwen3-14B (GRPO: B+C+LP)	19.59	**0.782**	**0.807**	**0.666**	0.231	**9.03**	**199.2**
	Qwen3-14B (GRPO: B+C+L+LP)	18.08	0.779	0.798	0.665	**0.234**	8.93	197.8
	Qwen3-14B (GRPO: B+C+S+LP)	**19.69**	0.778	0.801	0.662	0.231	9.02	193.3

Table 5: Model Performance Comparison for EN-DE Translation

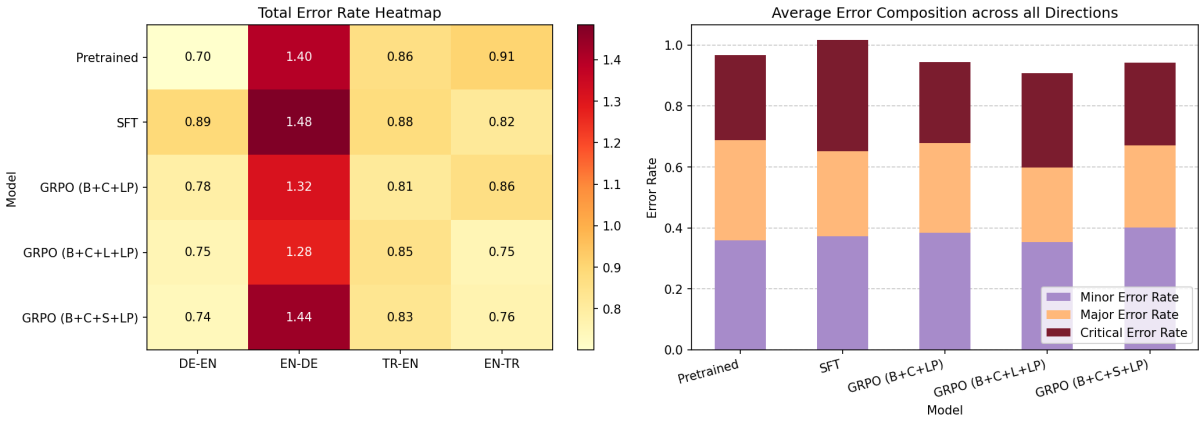

Figure 1: Error analysis with XCOMET on pretrained, SFT and GRPO trained Qwen3-14B. Left: Heatmap of total error rates (minor + major + critical) for each language direction. Right: Stacked bar chart of minor, major and critical error rates averaged over the four language directions.

The average error rates demonstrate that while SFT leads to a slight increase in total error rates compared to the pretrained baseline (primarily driven by a higher density of critical errors), the application of GRPO effectively reverses this trend. Specifically, the GRPO variants, led by the GRPO (B+C+L+LP) configuration with the LOO diversity reward, achieve the lowest overall error rates by significantly decreasing the frequency of major errors compared to the pretrained model and reducing critical errors relative to the SFT model. This suggests that our proposed GRPO framework with multi-objective reward functions also has the advantage of reducing the number of translation errors produced by pre-trained or fine-tuned Qwen3-14B.

5.2 Ablation Study

We conduct ablation studies on the English–Turkish direction, where we choose the Qwen3-14B (GRPO: B+C+S+LP) with the Self-BLEU diversity reward as a baseline since it yields the best balance between lexical richness and translation quality.

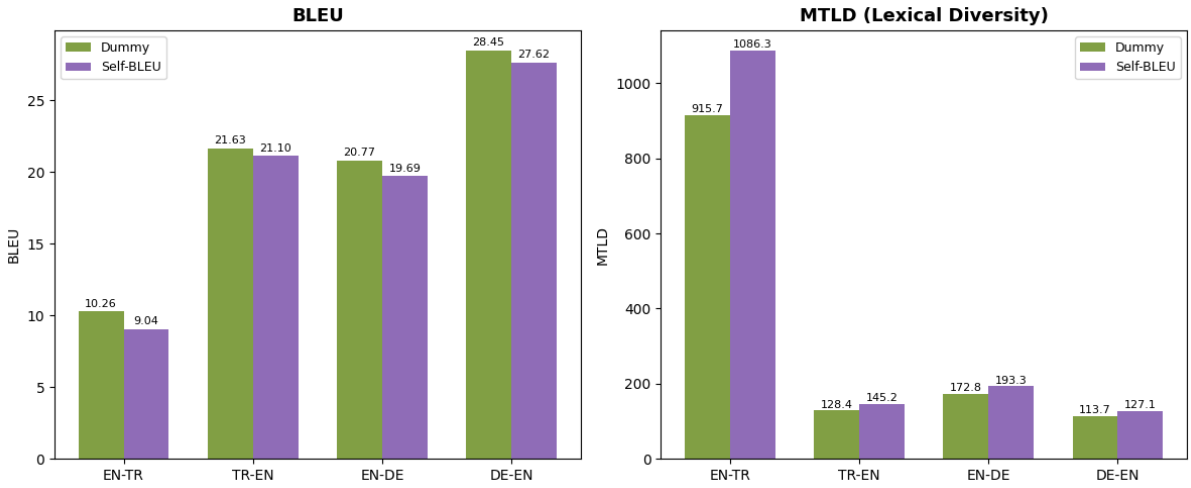

Figure 2: Comparison of GRPO runs with Self-BLEU diversity reward and with Dummy reward

Impact of Diversity Reward To isolate the specific impact of our proposed Self-BLEU diversity reward, we replace the diversity component with a zero-value dummy reward while keeping its weight ($W = 0.2$) and all other hyperparameters intact. This ensures the relative priority of quality and structural rewards remains constant. We GRPO-train Qwen3-14B with dummy setting on the same training sets for all language directions.

As shown in Figure 2, the most significant trend is the consistent increase in lexical diversity with the Self-BLEU reward. Across all language directions, the model trained with the diversity reward produced translations with higher lexical variance. This gain in diversity comes at a slight cost to BLEU scores. For example, in DE-EN, MTLD rose from 113.7 to 127.1, accompanied by a slight BLEU decrease from 28.45 to 27.62. Semantic metrics (COMET, BLEURT, COMET-Kiwi) showed negligible changes (± 0.001), which are detailed in Appendix D (Table 9). These results confirm that the diversity reward fosters lexical and syntactic exploration without sacrificing semantic faithfulness, proving superior for applications prioritizing diverse word choices.

Impact of KL Divergence Parameter To investigate the impact of the KL divergence parameter on model performance, we conducted an ablation study by varying the β parameter from 0 to 0.5 for the English–Turkish direction. We take (GRPO: B+C+S+LP) with the Self-BLEU diversity reward and $\beta = 0$ as baseline. As illustrated in Figure 3 (detailed in Appendix D, Table 10), the baseline configuration ($\beta = 0$) generally yields the most

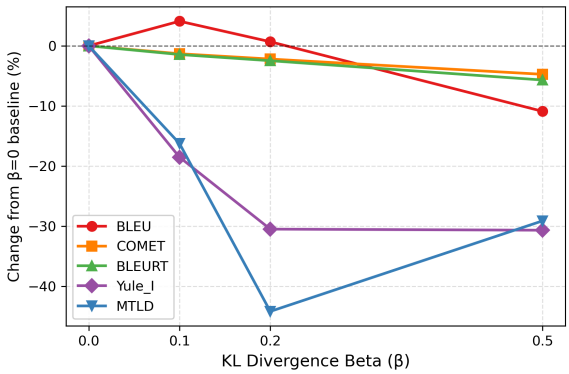

Figure 3: Effect of KL Divergence Parameter (β): Percent Change from the Baseline

robust performance across both neural and lexical metrics. We observe that while a small penalty ($\beta = 0.1$) provides a marginal improvement in BLEU scores (increasing from 9.04 to 9.41), neural metrics such as COMET and BLEURT exhibit a steady, monotonic decline as β increases. This suggests that enforcing a stricter proximity to the reference model distribution with a higher KL penalty constrains the model's ability to optimize for the reward signals provided by our proposed GRPO framework.

The most significant impact of the KL divergence penalty is observed in lexical diversity metrics. As shown in the percentage change plot, Yule's I and MTLD experience sharp degradations as β increases, with MTLD dropping by over 40% at $\beta = 0.2$ (from 1086.32 to 606.21). These results indicate that a high KL penalty severely stifles the model's linguistic variety, forcing it into more conservative, repetitive, or safe outputs that

Experiment	Weights	Quality Metrics				Diversity Metrics		
	(B/ C / S / LP)	BLEU	COMET	Kiwi	BLEURT	TTR	Yule's I	MTLD
Primary Baseline†	0.3 / 0.4 / 0.2 / 0.1	9.04	0.805	0.807	**0.632**	0.457	74.25	**1086.3**
Low Diversity	0.4 / 0.4 / 0.1 / 0.1	9.30	0.804	**0.811**	0.631	0.432	62.01	133.3
Balanced Variety	0.3 / 0.3 / 0.3 / 0.1	7.30	0.801	0.794	0.625	0.456	75.98	182.0
Diversity Stress Test	0.15 / 0.15 / 0.6 / 0.1	3.34	0.766	0.720	0.544	0.445	75.85	958.2
Neural Synergy	0.1 / 0.5 / 0.3 / 0.1	5.72	**0.807**	0.790	0.616	0.451	**85.74**	1083.1
Lexical Synergy	0.5 / 0.1 / 0.3 / 0.1	**9.74**	0.794	0.800	0.621	**0.462**	77.83	1034.7

Table 6: Ablation study of GRPO reward weights for EN-TR translation. Primary Baseline (†) represents the configuration used for main results. Weights are listed as ($R_{\text{BLEU}}/R_{\text{COMET}}/R_{\text{Self-BLEU}}/R_{\text{len}}$). Kiwi refers to COMET-Kiwi.

align closely with the base model but lack the richness of the unconstrained RL-tuned version. Consequently, for the EN-TR direction, a minimal or zero KL penalty appears optimal for maintaining both translation quality and lexical richness.

Reward Weight Sensitivity To investigate the trade-off between quality (R_{BLEU}, R_{COMET}) and diversity ($R_{\text{Self-BLEU}}$), we performed a reward weight sensitivity analysis on EN-TR with R_{len} fixed at 0.1 (Table 6). The *Primary Baseline* provides the most stable equilibrium between semantic accuracy and linguistic richness; therefore, it was chosen for our main experiments. In contrast, *Low Diversity* fails to improve lexical richness, resulting in an MTLD lower than the SFT-only model (Table 3).

The *Diversity Stress Test* serves as a warning against reward hacking, where the model generates diverse Turkish words with very low n-gram overlap to satisfy the diversity objective (0.6 weight), leading to a collapse in BLEU. However, the *Synergy* configurations demonstrate that anchoring diversity with strong quality rewards is effective. *Neural Synergy* reached the highest COMET and Yule's I scores, while *Lexical Synergy* achieves peak BLEU alongside high MTLD. These results confirm that $R_{\text{Self-BLEU}}$ effectively steers the model away from repetitive patterns without compromising lexical or semantic precision.

6 Conclusion

In this paper, we explored the impact of Group Relative Policy Optimization (GRPO) on lexical diversity and translation quality of large language models. Our primary contribution is the development of a multi-objective reward framework that explicitly incorporates diversity metrics.

Our experiments on Turkish–English and German–English demonstrate that our diversity-aware approach successfully enhances lexical richness alongside translation quality. Qwen3-14B models trained with our framework achieved state-of-the-art open-source translation quality, significantly reduced critical translation errors, and matched or surpassed the translation quality and the lexical richness of proprietary baselines like GPT-5.2. Ablation studies confirmed that explicit diversity rewards increase lexical variance without degrading semantic fidelity, and that a minimal KL divergence penalty is optimal for maintaining this richness. These results suggest that diversity-aware RL can be a powerful tool for optimizing LLMs towards more creative translation, particularly when the target domain demands linguistic variety and creativity.

For future work, we plan to explore other creativity measures, such as translation literalness and idiom translation. We also want to experiment with reasoning to observe its effect on creativity and quality. We will also conduct human evaluations to further assess the perceived creativity and stylistic sophistication of our produced translations.

7 Limitations

Due to the high computational cost of GRPO, we could not perform multiple independent runs. We report results from single, converged runs; however, stable reward trajectories suggest our performance gains are robust. To ensure reproducibility, we release our code and hyperparameters for future validation.

While base model exposure to public OPUS data is possible, we focus on the relative performance delta between GRPO and our baselines rather than absolute benchmark scores. The consistent improvements across models suggest that the optimization framework is effective regardless of the base model's prior exposure to test data.

Acknowledgments

This work was supported by Boğaziçi University Research Fund Grant Number 19986. The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

References

Anschel, Oron, Alon Shoshan, Adam Botach, Shunit Haviv Hakimi, Asaf Gendler, Emanuel Ben Baruch, Nadav Bhonker, Igor Kviatkovsky, Manoj Aggarwal, and Gerard Medioni. 2025. Group-aware reinforcement learning for output diversity in large language models. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32394–32415, Suzhou, China, November. Association for Computational Linguistics.

Chen, Xiwen, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Hao Wang, Haiyu Wu, Huayu Li, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. 2026. Dra-grpo: Your grpo needs to know diverse reasoning paths for mathematical reasoning.

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

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

Garcia Gilabert, Javier, Carlos Escolano, Xixian Liao, and Maite Melero. 2025. Terminology-constrained translation from monolingual data using GRPO. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 1335–1343, Suzhou, China, November. Association for Computational Linguistics.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models.

Guerberof-Arenas, Ana and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212, March.

He, Minggui, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, Hao Yang, Boxing Chen, and Osamu Yoshie. 2025. R1-t1: Fully incentivizing translation capability in llms via reasoning learning.

Hu, Jian, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization.

Ichihara, Yuki and Yuu Jinnai. 2025. Auto-weighted group relative preference optimization for multi-objective text generation tasks. In Potdar, Saloni, Lina Rojas-Barahona, and Sebastien Montella, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1134–1147, Suzhou (China), November. Association for Computational Linguistics.

Ichihara, Yuki, Yuu Jinnai, Tetsuro Morimura, Mitsuki Sakamoto, Ryota Mitsuhashi, and Eiji Uchibe. 2025. Mo-grpo: Mitigating reward hacking of group relative policy optimization on multi-objective problems.

Kalajdzievski, Damjan. 2023. A rank stabilization scaling factor for fine-tuning with lora.

Lannelongue, Loïc, Jason Grealey, and Michael Inouye. 2020. Green algorithms: Quantifying the carbon footprint of computation.

Li, Jiwei, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Knight, Kevin, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June. Association for Computational Linguistics.

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

Li, Chenyi, Yuan Zhang, Bo Wang, Guoqing Ma, Wei Tang, Haoyang Huang, and Nan Duan. 2026. Setpo: Set-level policy optimization for diversity-preserving llm reasoning.

McCarthy, Philip M. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-08-04.

Meng, Yu, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward.

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

Rei, Ricardo, 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, November. Association for Computational Linguistics.

Scalena, Daniel, Gabriele Sarti, Arianna Bisazza, Elisabetta Fersini, and Malvina Nissim. 2026. Steering large language models for machine translation personalization. In Demberg, Vera, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4681–4701, Rabat, Morocco, March. Association for Computational Linguistics.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Shao, Zhihong, 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.

Shen, Lucas. 2022. LexicalRichness: A small module to compute textual lexical richness.

Templin, Mildred C. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, volume 26. University of Minnesota Press, ned - new edition edition.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Uhlig, Kaden, Joern Wuebker, Raphael Reinauer, and John Denero. 2025. Cross-lingual human-preference alignment for neural machine translation with direct quality optimization. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 31–51, Suzhou, China, November. Association for Computational Linguistics.

Voigt, Rob and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In Elson, David, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz, editors, *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada, June. Association for Computational Linguistics.

Wei, Kangda and Ruihong Huang. 2026. Mmr-grpo: Accelerating grpo-style training through diversity-aware reward reweighting.

Wu, Minghao, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Xu, Haoran, 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.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, et al. 2025a. Qwen3 technical report.

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

Yirmibeşoğlu, Zeynep, Olgun Dursun, Harun Dalli, Mehmet Şahin, Ena Hodzik, Sabri Gürses, and Tunga Güngör. 2023. Incorporating human translator style into English-Turkish literary machine translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 419–428, Tampere, Finland, June. European Association for Machine Translation.

Yule, G.U. 1944. *The Statistical Study of Literary Vocabulary*. The University Press.

Zhu, Yaoming, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models.

A Dataset Statistics

In this section, we provide the detailed statistics for the literary translation datasets used in our experiments. Table 7 summarizes the number of sentences and token counts for each language pair and data split.

B Prompt Template and Chat Format

To ensure the model functions as a dedicated translation engine without conversational filler, we wrap the output in specific XML-style tags. The system message is designed to be highly restrictive:

> *"You are a strict [Source]-to-[Target] translation engine. Output only the [Target] translation of the user's text. Do not provide notes, explanations, or conversational filler. Do not explain slang."*

	TR-EN			DE-EN		
Split	**Sents**	**Tokens (tr)**	**Tokens (en)**	**Sents**	**Tokens (de)**	**Tokens (en)**
Train (SFT)	50,000	1,519,321	1,007,793	30,000	1,176,049	795,374
Train (GRPO)	5,000	153,528	101,904	5,000	194,495	131,702
Validation	500	16,105	10,726	500	19,836	13,198
Test	1,082	32,227	21,209	1,060	41,892	27,954

Table 7: Detailed statistics for the literary translation corpora. Token counts are calculated using the native Qwen3-14B tokenizer on the parallel text (excluding the prompt).

The chat template used for Qwen3-14B, which incorporates the `<translation>` tags and Jinja2 logic, is defined as follows:

Listing 1: Jinja2 Chat Template

```
{% for message in messages %}
   {% if message['role'] == 'system' %}
      {{ '<|im_start|>system\n' +
          message['content'] + '<|im_end
          |>\n' }}
   {% elif message['role'] == 'user' %}
      {{ '<|im_start|>user\n' + message
          ['content'] + '<|im_end|>\n' }}
   {% elif message['role'] == 'assistant
       ' %}
      {{ '<|im_start|>assistant\n<
          translation>' + message['
          content'] + '</translation><|
          im_end|>\n' }}
   {% endif %}
{% endfor %}
{% if add_generation_prompt %}
   {{ '<|im_start|>assistant\n<
       translation>' }}
{% endif %}
```

An example of the resulting sequence during inference for the **EN-TR** pair is provided below to illustrate the strict formatting requirements:

Listing 2: Example Inference Format

```
<|im_start|>system
You are a strict English-to-German
    translation engine. Output only the
    Turkish translation of the user's
    text. Do not provide notes,
    explanations, or conversational
    filler. Do not explain slang.
<|im_end|>
<|im_start|>user
You see that they are in pursuit of you,
     and that I am not lying to you.
<|im_end|>
<|im_start|>assistant
<translation> Sie sehen, dass sie Ihnen
    folgen, und dass ich Ihnen nicht
    etwas Unwahres gesagt habe.</
    translation>
<|im_end|>
```

For consistency across training stages, we use the same system prompt and chat template for both SFT and GRPO. The SFT phase serves to ensure the model strictly adheres to the desired output format, specifically the use of `<translation>` tags.

C Training Hyperparameters

Training for the SFT and GRPO stages is conducted for one epoch using 8-bit AdamW and sequence packing to optimize throughput. All training stages are performed on a single NVIDIA A100 or H100 GPU. Detailed hyperparameters, including learning rates and LoRA configurations are provided in Table 8. We present the sustainability statement of our experiments in Appendix E.

Parameter	**SFT**	**GRPO**
Learning Rate	5×10^{-5}	5×10^{-6}
Scheduler	Linear	Linear
Warmup Ratio	0.1	–
Optimizer	8-bit AdamW	8-bit AdamW
Batch Size	16	16
LoRA Rank (r)	64	64
LoRA Alpha (α)	128	128
Group Size (G)	–	8
Temperature	–	0.6
Max Length	512	512
KL β	–	0

Table 8: Hyperparameters for SFT and GRPO phases.

D Ablation Study Tables

We present the complete numerical results for our ablation studies in this section. Table 9 details the comparison between the dummy reward and the Self-BLEU diversity reward, while Table 10 shows the impact of the KL divergence penalty coefficient β on model performance.

E Sustainability Statement

Our experiments with 14B parameter models run in 53.8h on 1 GPU NVIDIA A100 and 188.1h

Dir.	Setting	BLEU	COMET	Kiwi	BLEURT	MTLD
EN-TR	Dummy Reward	10.26	0.805	0.814	0.631	915.70
	Self-BLEU	9.04	0.805	0.807	0.632	1086.32
TR-EN	Dummy Reward	21.63	0.806	0.818	0.649	128.44
	Self-BLEU	21.10	0.806	0.820	0.648	145.24
DE-EN	Dummy Reward	28.45	0.792	0.789	0.658	113.71
	Self-BLEU	27.62	0.789	0.791	0.656	127.12
EN-DE	Dummy Reward	20.77	0.777	0.801	0.658	172.79
	Self-BLEU	19.69	0.778	0.801	0.662	193.31

Table 9: Complete Ablation Study Results: Dummy Reward vs. Self-BLEU Diversity Reward. 'Dummy' refers to setting the diversity reward weight to 0. Kiwi refers to COMET-Kiwi.

β	BLEU	COMET	COMET-kiwi	BLEURT	TTR	Yule's I	MTLD
0.0	9.04	0.8046	0.8068	0.6319	0.4572	74.25	1086.32
0.1	9.41	0.7940	0.8086	0.6227	0.4493	60.49	909.70
0.2	9.10	0.7869	0.8063	0.6163	0.4382	51.60	606.21
0.5	8.05	0.7665	0.7951	0.5959	0.4365	51.47	769.92

Table 10: Effect of the KL divergence penalty coefficient β on translation quality (BLEU, COMET, COMET-kiwi, BLEURT) and lexical diversity (TTR, Yule's I, MTLD) for EN→TR.

on 1 GPU NVIDIA H100, and draw a total of 183.82 kWh. Based in a regional high-performance computing center, this has a carbon footprint of 84.56 kg CO_2e, which is equivalent to 7.69 tree-years (Lannelongue et al., 2020). To optimize throughput and minimize energy usage, all training stages—including supervised fine-tuning and Group Relative Policy Optimization (GRPO)—utilized 8-bit AdamW and sequence packing.