Reinforcement Learning for Machine Translation via Structural, Lexical, and Semantic Rewards

Oleg Dats

June 18, 2025

Abstract

We propose a novel approach to training neural machine translation (NMT) models using reinforcement learning (RL) guided by a diverse set of interpretable, reference-free reward functions. These rewards capture structural, lexical, and semantic fidelity between the source and translated text—measuring aspects such as punctuation consistency, word and POS count alignment, lexical diversity, and preservation of formatting (e.g., uppercase letters and symbols). In addition, we incorporate semantic rewards based on COMET-Kiwi and cosine similarity of sentence embeddings. This framework enables translation models to be fine-tuned without requiring parallel reference texts and encourages translations that are syntactically balanced, semantically accurate, and structurally faithful to the source.

1 Introduction

Challenge: Existing RL-based MT optimization strategies (e.g., COMET, BLEU) often bias toward semantic preservation while ignoring syntactic or lexical fidelity, leading to generic outputs.

Insight: Structural and lexical constraints can regularize policy optimization, avoiding degenerate behaviors.

Our Contributions:

- We design a reward function composed of structural, lexical, and semantic components.
- We fine-tune a Gemma 4B model using GRPO on English-only WikiPar with reference-free rewards.
- Evaluations on FLORES++ using GPT-4o LLM-as-a-judge show state-of-the-art fluency and fidelity for Ukrainian.

Importance of Reference-Free Translation Training

Reference-free translation training is essential for language pairs with limited or low-quality parallel corpora, such as English–Ukrainian. Using source-side monolingual data and interpretable reference-free metrics enables scalable training for morphologically rich, low-resource languages.

2 Method

2.1 Role of Rewards

In our reinforcement learning framework for NMT, structural and lexical rewards play a crucial role in guiding the model toward well-formed, interpretable translations, rather than exploiting shortcuts that maximize semantic scores alone. Without these auxiliary rewards, the model may learn to "hack" semantic metrics—e.g., by generating overly generic or semantically similar outputs that lack grammatical structure, proper formatting, or lexical richness. By enforcing constraints on syntax, formatting, and word distribution, structural and lexical rewards anchor the training signal, ensuring that high semantic scores correspond to translations that are also linguistically and structurally faithful to the source.

2.2 Reward Modeling

We define the total reward as:

$$r_{\text{total}} = \lambda_{\text{struct}} \cdot r_{\text{struct}} + \lambda_{\text{lex}} \cdot r_{\text{lex}} + \lambda_{\text{sem}} \cdot r_{\text{sem}}$$

where each reward component is normalized, and the weights λ are set empirically.

Structural Rewards

1. **Uppercase Letter Count Alignment**
   Measures the relative difference in the number of uppercase letters between the source and translation. Encourages the preservation of proper nouns, acronyms, and emphasis.

2. **Non-Letter Character Count Alignment**
   Compares the count of punctuation marks, digits, and other special characters (excluding letters and spaces) between source and translation. Promotes fidelity in formatting and numeric/symbolic information.

3. **Word Count Alignment**
   Rewards translations that have a similar total number of words to the source. Helps enforce overall length consistency.

4. **Unique Word Count Alignment**
   Compares the number of unique words in the source and translation, encouraging lexical variety and discouraging over-repetition.

5. **Unpaired Symbol Count**
   Compares the sets of special symbols between the source and translation to identify unmatched characters. Rewards translations that preserve similar symbol distributions (e.g., digits, punctuation).

Lexical Rewards

6. **POS Count Alignment**
   Compares the counts of major part-of-speech categories (NOUN, VERB, ADJ) between the source and translation. Encourages structural similarity at the grammatical level, maintaining linguistic balance and expressiveness.

Semantic Rewards

7. **COMET-Kiwi Quality Score**
   Uses the COMET-Kiwi model to score translation quality without reference. Rewards outputs that are semantically and fluently close to the source sentence.

8. **Embedding Cosine Similarity**
   Computes cosine similarity between multilingual sentence embeddings (e.g., via GTE-multilingual). Encourages semantic preservation by aligning the meaning of the translation to the source.

2.3 Reinforcement Learning Framework

We use the GRPO algorithm without reasoning tokens. For each source sentence, we sample n translations, compute the reward for each, and update the policy using GRPO.

3 Experiments

3.1 Datasets

Train: WikiPar (English monolingual)
Test: FLORES++ (English–Ukrainian subset)

3.2 Evaluation

- **GEA5, GEA100:** GPT-4o-based stylistic evaluations on 5- and 100-point scales, focusing on fluency, coherence, and literary quality.

- **GRF:** GPT-4o-based reference-free score assessing semantic fidelity on a 0–100 scale.
- **Internal Metrics:** COMET-Kiwi (semantic quality), Structural score (format preservation), Lexical alignment (POS balance).

3.3 Baselines

- Supervised Gemma 4B (SFT)
- RL with only COMET-Kiwi
- RL with semantic + structural rewards
- RL with all rewards (ours)

3.4 Results (Mock)

Dataset	Model	GEA5	GEA100	GRF	Struct.	Lex.	COMET
WikiPar (validation)							
	Gemma-SFT	3.2	66.5	83.7	0.71	0.66	0.60
	RL (sem)	3.4	70.2	85.9	0.59	0.60	0.64
	RL (sem+struct)	3.7	73.8	87.5	0.83	0.73	0.68
	RL (all)	**4.1**	**77.6**	**88.9**	**0.88**	**0.79**	**0.71**
FLORES++ (test)							
	Gemma-SFT	3.4	68.2	85.1	0.72	0.68	0.61
	RL (sem)	3.6	71.5	86.7	0.60	0.61	0.66
	RL (sem+struct)	3.9	75.1	88.0	0.84	0.75	0.69
	RL (all)	**4.3**	**78.9**	**89.4**	**0.89**	**0.81**	**0.72**

Table 1: Translation quality on WikiPar (validation) and FLORES++ (test) across different reward configurations.

3.5 Case Studies

We present qualitative improvements in formatting, name translation, and POS alignment.

4 Analysis & Ablations

- Dropping lexical rewards leads to repetition.
- Removing structural rewards reduces punctuation fidelity.

5 Related Work

DeepTrans introduced reward modeling in deep reasoning LLMs for literature translation.
Prior RL-NMT methods focused on BLEU/COMET (Ranzato et al., Shen et al.).
LLM-as-a-judge is increasingly used for reference-free evaluation (e.g., GPT-4o, Kocmi et al.).

6 Conclusion

We demonstrate that combining structural, lexical, and semantic reference-free rewards in RL yields high-quality English–Ukrainian translations without requiring target-side supervision. This framework is promising for low-resource, morphologically rich language pairs.

7 Preliminary Results (Gemma 1b)

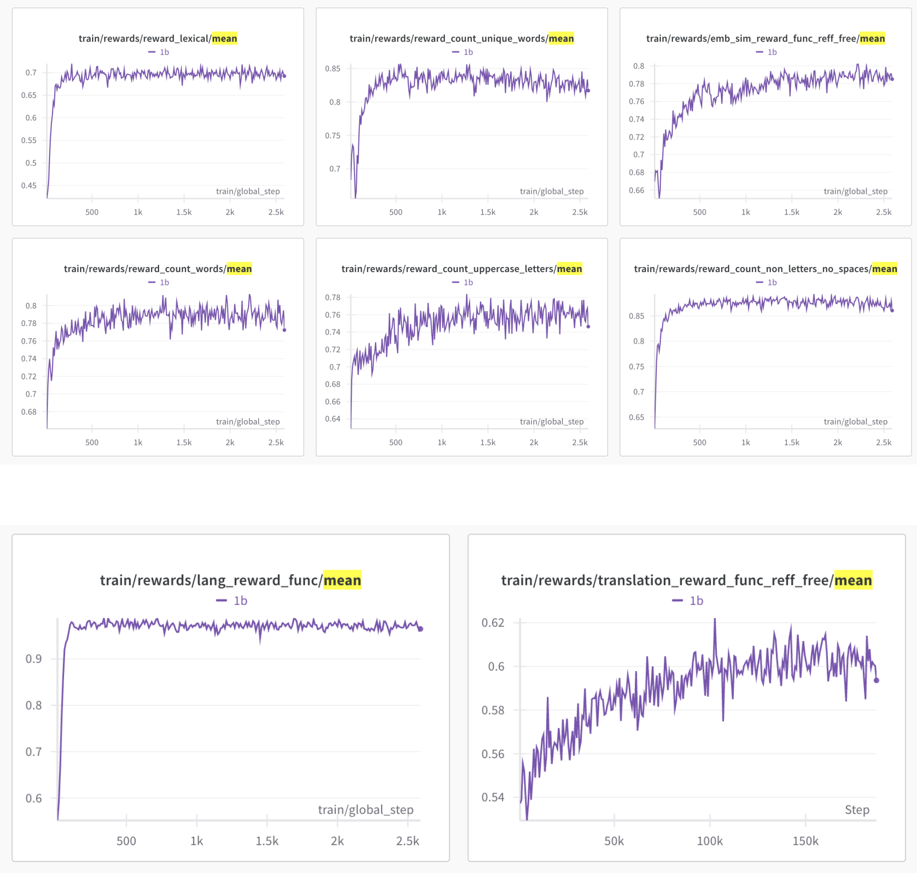

Figure 1: Train reward components: structural, lexical, semantic (reward shaping curves).

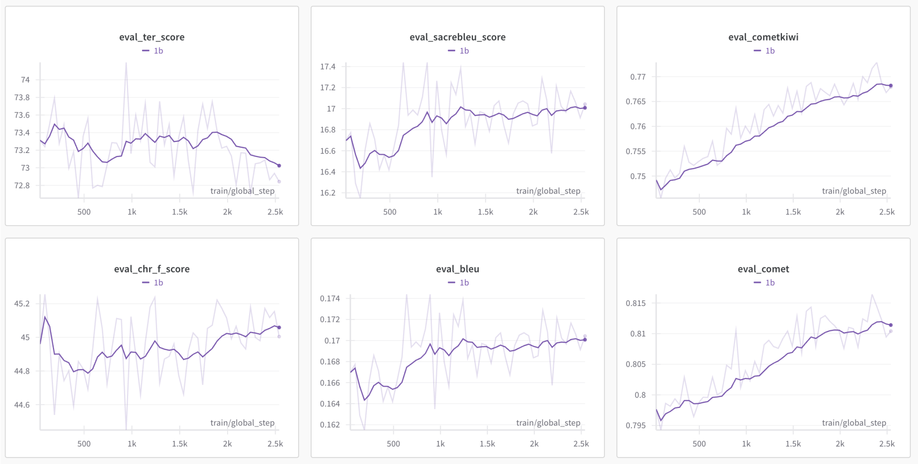

Figure 2: Evaluation metrics: BLEU, COMET, chrF, TER, SacreBLEU, and COMET-Kiwi over steps.