Fine-Grained Reward Optimization for Machine Translation using Error Severity Mappings

Miguel Moura Ramos[*1,2] **Tomás Almeida**[1] **Daniel Vareta**[1] **Filipe Azevedo**[1,2]
Sweta Agrawal[2] **Patrick Fernandes**[1,2,3] **André F. T. Martins**[†1,2,4]

[1]Instituto Superior Técnico, Universidade de Lisboa (ELLIS Unit Lisbon), Portugal
[2]Instituto de Telecomunicações, Portugal [3]Carnegie Mellon University, USA
[4]TransPerfect, Portugal

Abstract

Reinforcement learning (RL) has been proven to be an effective and robust method for training neural machine translation systems, especially when paired with powerful *reward models* that accurately assess translation quality. However, most research has focused on RL methods that use sentence-level feedback, leading to inefficient learning signals due to the *reward sparsity* problem—the model receives a single score for the entire sentence. To address this, we propose a novel approach that leverages fine-grained, token-level quality assessments along with error severity levels using RL methods. Specifically, we use xCOMET, a state-of-the-art quality estimation system, as our token-level reward model. We conduct experiments on small and large translation datasets with standard encoder-decoder and large language models-based machine translation systems, comparing the impact of sentence-level versus fine-grained reward signals on translation quality. Our results show that training with token-level rewards improves translation quality across language pairs over baselines according to both automatic and human evaluation. Furthermore, token-level reward optimization improves training stability, evidenced by a steady increase in mean rewards over training epochs.

1 Introduction

Neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014), a leading approach within MT, leverages neural networks to automate language translation and has driven significant improvements in translation quality. However, most NMT systems are predominantly trained using *maximum likelihood estimation* (MLE). MLE-based training focuses on maximizing the probability of next-word predictions given a partial reference. This often leads to a critical problem known as exposure bias—the model uses ground-truth prefix tokens during training, but during inference it relies on its previous predictions (Bengio et al., 2015; Ranzato et al., 2016; Wiseman and Rush, 2016). This can cause errors to propagate through the generated sequence, severely degrading the translation quality. Furthermore, it tends to produce translations that lack global coherence and adequacy as the model does not sufficiently consider the context of entire sentences or the overarching meaning. This has spurred interest in using alternative approaches that leverage RL methods for training NMT systems.

RL-based approaches use explicit reward models to evaluate the outputs generated by the NMT system, assigning scores to generated hypotheses to guide the learning process. However, most prior research (Ranzato et al., 2016; Wu et al., 2016; Bahdanau et al., 2017; Nguyen et al., 2017; Wu et al., 2017; Kreutzer et al., 2018a,b; Kiegeland and Kreutzer, 2021) predominantly relies on sentence-level feedback and often struggles with *reward sparsity*, particularly for long-form text generation: Sentence-level rewards fail to capture specific issues within a translation, making it difficult for the model to learn from negative reward signals. As shown in Figure 1, two translations corresponding to different source texts of varying length receive the same sentence-level quality score of 70, yet differ significantly in the nature and impact of the errors: The first translation has several minor errors scattered throughout the text, while the latter has major errors that could potentially hinder the understanding of the

[*]Core contributor and corresponding author: miguel.moura.ramos@tecnico.ulisboa.pt.
[†]Work done while at Unbabel.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 733–754, 2026. https://doi.org/10.1162/TACL.a.646
Action Editor: Giuseppe Carenini. Submission batch: 2/2025; Revision batch: 9/2025; Published 5/2026.
© 2026 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

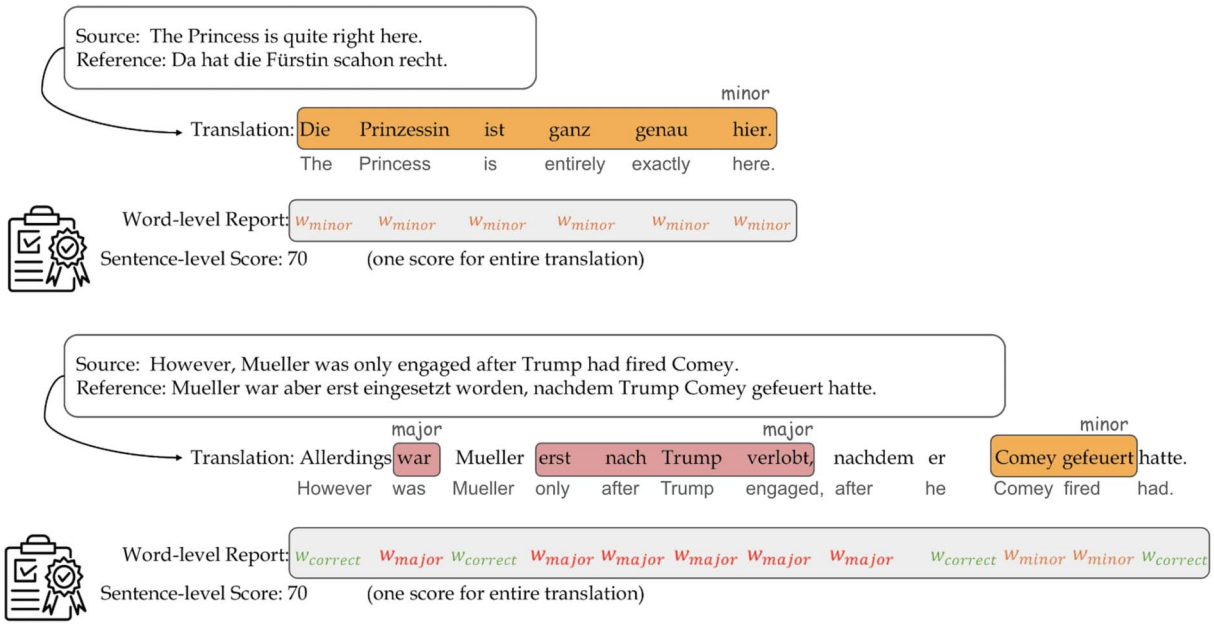

Figure 1: Two examples are presented, both with identical sentence-level assessments but differing error severity and frequency. The reward model identifies translation error spans along with their corresponding severity levels. In these examples, we highlight both minor and major error spans. By mapping these spans to numerical values that reflect their severity, we can derive word-level scores/rewards. Since error spans can contain multiple words, we assume that all words within a given span share the same severity.

original content. This suggests that learning can be more effective if feedback is provided at a fine-grained level, including precise identification of the nature of errors.

Recent advancements in automated MT evaluation metrics that generate fine-grained error span predictions, such as xCOMET (Guerreiro et al., 2024), MetricX (Juraska et al., 2023), AutoMQM (Fernandes et al., 2023), EAPrompt (Lu et al., 2024), MaTESE (Perrella et al., 2022), and BARTScore++ (Lu et al., 2023) have shown promise in improving alignment with human translation quality judgments. These metrics directly predict token-level error severity (no error, minor, major, or critical) and optionally provide sentence-level quality assessments or prompt large language models to identify error types (e.g., mistranslation, omission) and severities based on the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014).

Despite the potential of severity-based metrics to improve translation quality, their application in MT training via RL methods remains relatively underexplored, since it presents several challenges: (i) the feedback, albeit informative and frequent, can be noisy, and (ii) determining the appropriate reward assignments for different severity levels to ensure effective and stable learning is not straightforward. In this regard, our research aims to answer the following questions:

1. Do fine-grained RL methods offer benefit over sentence-level feedback in improving translation quality and stabilizing training?

2. Can fine-grained MT metrics be effectively used to provide accurate, detailed, human-aligned feedback to reduce reward sparsity?

When answering these questions, we make the following contributions:

1. We propose using a fine-grained evaluation metric, xCOMET, to generate token-level rewards, which increases the reward density by providing frequent token-level rewards, thus improving the robustness and stability of RL-based MT.

2. We introduce a new severity map to effectively use the reward signals, overcoming the limitations of standard MQM scoring, as demonstrated in our experimental results.

3. We conduct experiments on English-to-German (EN→DE), English-to-French

(EN→FR), German-to-English (DE→EN), and French-to-English (FR→EN) translation datasets, comparing the overall translation quality of NMT systems when using sentence and token-level rewards, showing that translation quality improves when employing xCOMET as a reward model.

By integrating fine-grained reward signals into NMT training, we demonstrate significant improvements in translation quality and overcome the challenges of exposure bias, reward sparsity, and instability of RL training, paving the way for more reliable and accurate MT systems.

2 Background

Standard NMT Training. NMT systems utilize learnable parameters, denoted as θ, to estimate the probability distribution $p_\theta(y|x)$ over a set of possible translations $\mathcal{Y}$, conditioned on a given source sentence x. In the simplest form of NMT training, *maximum likelihood estimation* (MLE) is used, which maximizes the probability of the correct target translation y given the source sentence x. The MLE objective can be expressed as:

$$\mathcal{L}_{\text{MLE}}(\theta) = \sum_{(x,y)\in D} \log p_\theta(y|x), \qquad (1)$$

where D represents a dataset of parallel sentences.

Limitations of MLE Training. While commonly used in NMT, MLE has several limitations, primarily its weak learning signals from token-level feedback. As MLE assumes gold-reference tokens (teacher-forcing) during training, when exposed to its own incorrect predictions during inference, it can lead to error accumulation and poor performance on longer sequences. Another major limitation is its tendency to optimize for a single ''most likely'' translation, often ignoring the variety of equally valid alternatives, which reduces the model's ability to generate diverse and natural outputs. Additionally, MLE is sensitive to noisy or inconsistent reference translations, which can degrade performance by producing unreliable gradient updates. Taken together, these challenges have prompted the exploration of RL methods, which offer more effective feedback on model-generated outputs by optimizing directly for downstream translation quality measures.

Formulating MT as an RL Problem. In the context of MT, we can model the translation process as a Markov Decision Process (MDP) (Puterman, 1990), defined by the tuple (S, A, P, R, γ) with a finite vocabulary $\mathcal{V}$. The state space S consists of all possible sequences of tokens up to the current time step, which includes the input sequence in the source language, as well as the target language tokens generated so far. Initially, the state s_0 corresponds to the input sentence in the source language, $x = (x_1, x_2, \ldots, x_l)$, where each token $x_i \in \mathcal{V}_{\text{source}}$. At each time step $t \in [0, T]$, the state s_t represents the sequence of tokens generated up to that point, which can be expressed as:

$$s_t = (x_1, x_2, \ldots, x_l, \hat{y}_0, \hat{y}_1, \ldots, \hat{y}_{t-1}).$$

The agent selects an action $\hat{y}_t \in A$, which is a token generated by the policy p_θ based on the current state s_t. The process continues until an end-of-sequence token is generated, completing the translation. The reference tokens in the target language are denoted by $y = (y_1, y_2, \ldots, y_m)$, where $y_t \in \mathcal{V}_{\text{target}}$. The generated tokens $\hat{y}_t$ are evaluated against y_t to measure the quality of the translation. For $t > 0$, the state transition function $P : S \times A \to [0, 1]$ defines the probability of transitioning from one state to another by appending a chosen token to the current translation, and the reward function $R : S \times A \to \mathbb{R}$ assigns a real-valued reward r to each transition $(s, \hat{y})$, where $s \in S$ and $\hat{y} \in A$, based on the quality of the generated translation sequence. Conceptually, the reward function is defined as a mapping from a hypothesis $\hat{y}$ to a score, i.e., $R(\hat{y})$. In practice, many MT metrics additionally condition on the source and/or the reference, which we make explicit as $R(x, \hat{y}, y)$. Formally, the reward function can be written as:

$$R(s_t, \hat{y}_t) = R(x, \hat{y}_{<t}, \hat{y}_t, y) = r.$$

Sentence-level rewards are provided only once per translation and evaluate the entire output at once. Token-level rewards, on the other hand, give feedback for each generated token. The discount factor $\gamma \in [0, 1]$ is used to weigh future rewards, with $\gamma = 1$ typically chosen in MT to ensure that all rewards are valued equally, allowing the optimization of the entire sequence of tokens in the translation rather than focusing on just the initial tokens. Finally, the goal is to maximize the expected cumulative reward over trajectories

$$\mathcal{L}_{\text{REINFORCE}}(\theta) = \mathbb{E}_{\hat{y}\sim p_\theta(y|x)}\left[R(\hat{y}) \log p_\theta(\hat{y}|x)\right] \quad (2)$$

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_{\hat{y}\sim p_\theta(y|x)}\left[\min\left\{\frac{p_\theta(\hat{y}|x)}{p_{\text{old}}(\hat{y}|x)}\hat{A}_{x,\hat{y}},\ \text{clip}\left(\frac{p_\theta(\hat{y}|x)}{p_{\text{old}}(\hat{y}|x)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{x,\hat{y}}\right\}\right] \quad (3)$$

Figure 2: Sentence-level RL losses.

$\hat{y}$ sampled from the p_θ. The objective function can be written as:

$$\mathcal{L}_{\text{RL}} = \mathbb{E}_{\hat{y}\sim p_\theta}\left[\sum_{t=0}^{T} R(x, \hat{y}_{<t}, \hat{y}_t, y)\right].$$

Policy Gradient Algorithms. As illustrated in Figure 2, we can optimize the above objective using policy gradient methods: REINFORCE (Williams, 1992; Ranzato et al., 2016), a vanilla policy gradient method, optimizes translation by sampling hypotheses $\hat{y} \sim p_\theta(y|x)$, scoring them with a reward obtained from an MT metric $R(\hat{y})$, and updating the model to maximize expected rewards, as shown in Equation 2. Despite its simplicity, it often struggles with high variance and instability. Proximal Policy Optimization (PPO) (Schulman et al., 2017) mitigates this by using a clipped surrogate objective (Equation 3) to keep policy updates stable within a margin ϵ, efficient and employs Generalized Advantage Estimation (GAE) (Schulman et al., 2016) to compute advantages $\hat{A}$ using rewards R and value function V. While PPO performs well across various tasks, simpler methods like REINFORCE can sometimes rival or surpass it (Ahmadian et al., 2024). Both are evaluated in our experiments (§5.2).

3 Related Work

Advancements and Challenges in Sentence-Level Feedback. Incorporating human feedback as rewards and optimizing language models with RL methods effectively aligns them with human preferences (Ouyang et al., 2022), often surpassing MLE. A notable example in translation tasks is Minimum Risk Training (MRT) (Shen et al., 2016), which minimizes expected risk based on evaluation metrics to directly improve translation quality. Recent advances in NMT build on this idea by refining training with feedback from metrics or human evaluations, incorporating alignment techniques and RL methods (Nguyen et al., 2017; Kreutzer et al., 2018b; Wu et al., 2018; Kiegeland and Kreutzer, 2021; Xu et al., 2024a; Agrawal et al., 2024; Zhu et al., 2024; He et al., 2024; Ramos et al., 2024). Despite these advancements, sentence-level feedback methods face persistent challenges such as sparse rewards, instability, and difficulty handling long sequences (Wu et al., 2018). These issues hinder performance, generalization, and robust learning, even with multi-objective optimizations (Wu et al., 2023; Jang et al., 2024). To address the limitations of sentence-level feedback, recent research has explored finer-grained rewards at the token level for tasks such as language model alignment (Xia et al., 2024; Yoon et al., 2024; Guo et al., 2024; Cao et al., 2024), controllable text generation (Li et al., 2024), query generation (Ouyang et al., 2024), among others, but remain relatively underexplored in MT.

Token-Level Feedback and Reward Modeling for MT. Previous approaches to token-level reward modeling often relied on binary error markings generated by humans (Kreutzer et al., 2020; Domingo et al., 2017) or simulated it by comparing model predictions with reference translations based on heuristic methods (Petrushkov et al., 2018). While effective, these methods provide limited feedback due to their binary nature and require costly human annotation, making them less practical for scalable solutions. Other approaches have employed reward shaping techniques (Ng et al., 1999; Wu et al., 2018; Goyal et al., 2019; Devidze et al., 2022), incorporating intermediate rewards along with BLEU (Papineni et al., 2002) as the reward function. However, partial BLEU or token-level BLEU are less effective for fine-grained reward modeling, as they depend on exact N-gram matching and fail to capture meaningful semantic differences and context. Consequently, these methods, while valuable, are limited in their granularity and fail to address the severity of errors introduced at the token level.

4 Approach

In this section, we present our method for incorporating token-level rewards into RL training for

machine translation (MT). To address the limitations of prior approaches, such as binary feedback or coarse sentence-level scores, we use token-level rewards derived from state-of-the-art evaluation metrics that predict error spans and severity levels. These fine-grained signals are then used to guide learning through adaptations of REINFORCE and PPO objectives at the token level, enabling more effective and stable training of MT systems.

Token-Level Reward Modeling. Building on the MDP formulation for MT, we focus on token-level reward modeling (feedback is provided for individual tokens rather than entire sequences) allowing the model to refine its policy by identifying and addressing specific translation errors. Given an evaluation metric, $\mathcal{M}$ that predicts error spans along with their severity levels (e.g., minor, major, critical) for a hypothesis given source and optionally a gold reference, we assign numerical weights to each token within an error span according to a severity mapping as defined below:

$$\text{SEVERITY MAP} = \begin{cases} \text{correct span} & : w_{\text{correct}}, \\ \text{minor error} & : w_{\text{minor}}, \\ \text{major error} & : w_{\text{major}}, \\ \text{critical error} & : w_{\text{critical}}. \end{cases}$$

We use the evaluation metric xCOMET as $\mathcal{M}$ as it was shown to achieve the best correlation with human judgments and was the winning submission for the WMT23 Metrics Shared Task (Freitag et al., 2023). The severity weights from xCOMET adhere to the MQM framework (Lommel et al., 2014), which classifies translation issues into categories such as fluency, adequacy, grammar, and style. Each token within an error span is assigned the same severity weight (see Figure 1). We note that although the weights follow the MQM guidelines, they need to be further adjusted depending on the tasks to optimize the performance of token-level RL.

Tokenization-Agnostic Reward Assignment. MT systems and evaluation models typically use subword-level tokenization methods such as Byte-Pair Encoding (Gage, 1994, BPE) or SentencePiece (Kudo and Richardson, 2018), where words can be split into multiple subword tokens and token boundaries may not align with natural word boundaries. Given a detokenized hypothesis from the MT system, our evaluation model $\mathcal{M}$ produces error spans defined at the character level. To assign rewards at the token level for the tokenized hypothesis $\hat{y}$, we first re-tokenize the detokenized hypothesis using the same tokenizer applied during model training. This allows us to obtain precise character offsets for each subword token. We then align tokens to the character-level error spans by checking for overlap: any token whose character span overlaps with an error span inherits the corresponding error severity. This alignment avoids relying on explicit word boundaries or whitespace segmentation, making the reward assignment robust to different tokenization schemes—including those that generate cross-word subword units—and applicable across languages with or without explicit word boundaries. By grounding token-level rewards in character-level overlap rather than word-based grouping, our method ensures consistency and generalizability across tokenization models and languages. Finally, if a token overlaps multiple spans, it is assigned the worst severity (critical $>$ major $>$ minor $>$ correct) avoiding averaging or length-weighting to remain tokenizer-agnostic. Formally, for a token t with overlaps $E(t)$, $\ell(t) = \max_{\succ}\{\ell(e) \mid e \in E(t)\}$; if $E(t)$ is empty, $\ell(t) =$ correct. This severity is then mapped to its numeric reward (Table 4). The full algorithm is provided in Appendix A.

Token-Level Policy Refinement. In token-level RL, we maintain the structure of traditional sentence-level RL losses but adjust them to operate at the token level. We generate the full sequence, compute rewards for each token, and then perform updates for each token separately. The traditional sentence-level REINFORCE objective (Equation 2) is adapted to token-level by adjusting the loss to calculate the reward for each individual token. After generating the full sequence, we perform updates for each token one at a time, as follows:

$$\mathcal{L}_{RL}(\theta) = \mathbb{E}_{\hat{y}\sim p_\theta(y|x)}\left[\sum_{t=0}^{T} R(x, \hat{y}_{<t+1}, y) \log p_\theta(\hat{y}_t|\hat{y}_{<t}, x)\right]. \tag{4}$$

Here, $R(x, \hat{y}_{<t+1}, y)$ is the reward for token $\hat{y}_t$, reflecting its contribution to the overall sequence. Similarly, we extend the sentence-level

PPO objective (Equation 3) to token-level by modifying the loss function to compute the policy ratio and advantage for each token independently. The token-level PPO objective is defined as:

$$\mathcal{L}_{RL}(\theta) = \mathbb{E}_{\hat{y}\sim p_\theta(y|x)}\left[\sum_{t=0}^{T} \min\left\{\frac{p_\theta(\hat{y}_t|\hat{y}_{<t},x)}{p_{\text{old}}(\hat{y}_t|\hat{y}_{<t},x)}\hat{A}_{x,\hat{y}_{<t}},\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{p_\theta(\hat{y}_t|\hat{y}_{<t},x)}{p_{\text{old}}(\hat{y}_t|\hat{y}_{<t},x)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{x,\hat{y}_{<t}}\right\}\right]. \quad (5)$$

The policy ratio captures the change in policy for each token $\hat{y}_t$ relative to the previous policy. The token-level advantage is estimated using Generalized Advantage Estimation (GAE) (Schulman et al., 2016), which balances bias and variance by mixing temporal-difference (TD) errors across multiple steps. The advantage at time step t is computed as:

$$A_t = \sum_{l=0}^{T-t-1} \lambda^l \delta_{t+l},$$

where $\delta_t = R(x, \hat{y}_{<t+1}, y) + V(x, \hat{y}_{<t+1}) - V(x, \hat{y}_{<t})$ is the temporal-difference (TD) error, $\lambda \in [0, 1]$ is the GAE parameter, r_t is the reward at time step t, and $V(x, \hat{y}_{<t})$ is a learned value function that estimates the expected return from the state defined by the input x and the generated prefix $\hat{y}_{<t}$. As explained earlier, we consider $\gamma = 1.0$ for our use case, and therefore omit the discount factor for notational simplicity. To ensure stable training, we apply clipping to limit the extent of policy updates, preventing large, unstable shifts. This approach allows for more granular control over the model's learning, ensuring that each token is generated in a way that maximizes task-specific objectives while maintaining stability in the policy updates. Clipping also helps mitigate length bias by preventing longer sequences from accumulating disproportionately high rewards.

5 Experiments

We outline the experiments designed to explore the application of RL for MT, specifically focusing on comparing the impact of sentence-level and token-level reward signals.

5.1 Experimental Setup

Models. We use three state-of-the-art models: a standard encoder-decoder MT model, NLLB (NLLB Team et al., 2024), and two LLM-based MT systems, TOWER (Alves et al., 2024) and GEMMA (Rivière et al., 2024). While NLLB and TOWER are dedicated MT models optimized for translation tasks, GEMMA is an LLM that exhibits strong multilingual capabilities. These models differ in both their architectures and pre-training methodologies. Each is pre-trained on diverse multilingual datasets, establishing them as robust baselines for investigating the effects of SFT and RL techniques.

Data. We use the following training datasets in our experiments: (1) The IWSLT2017 dataset (Cettolo et al., 2017), with 242k examples for English-French (EN↔FR), supports rapid experimentation and frequent training iterations. (2) The WMT18 dataset (Bojar et al., 2018) contains 42.3M examples for English-German (EN↔DE). We train NLLB with both datasets and the LLM-based models with (2). Training stops once rewards stabilize, so not all examples are used. We evaluate NLLB models using their respective test splits: IWSLT17 (EN↔FR) and WMT18 (EN↔DE). To standardize comparison across MT systems (NLLB and LLM-based MT systems), we also evaluate all models on the WMT24 dataset (Kocmi et al., 2024), addressing concerns about data contamination, as TOWER training included the WMT18 test set.

Evaluation. We assess translation quality using a comprehensive suite of well-established evaluation metrics. These include **lexical reference-based metrics**, such as BLEU (Papineni et al., 2002) and CHRF (Popović, 2015); **neural reference-based metrics**, including COMET22 (Rei et al., 2022), xCOMET (Guerreiro et al., 2024), and BLEURT (Sellam et al., 2020); and a **neural reference-free metric**, COMETKIWI-23 (Rei et al., 2023). Lexical metrics focus on word overlap and N-gram matching, while neural metrics evaluate translations in terms of semantic coherence and contextual quality. Including a reference-free metric enables evaluation without reliance on predefined reference texts. This diverse set of metrics captures multiple dimensions of translation quality, including fluency, grammatical accuracy, semantic adequacy, and contextual relevance. By using various evaluation criteria, we reduce potential biases that may arise from aligning the reward model with a single evaluation metric, ensuring

more robust and reliable conclusions about the impact of different approaches on translation quality.

We apply significance testing at a confidence threshold of 95%. For segment-level metrics, such as COMET-22, we test at the segment level, but for corpus-level metrics, such as BLEU and ChrF, we apply bootstrapping with 100 samples of size 500 (Koehn, 2004). Performance clusters are formed based on statistically significant gaps, and final rankings are derived by averaging the cluster scores across all languages (Colombo et al., 2022; Freitag et al., 2023). In addition to automated metrics, we conduct human evaluations with two professional annotators, reporting inter-annotator agreement (Pearson's r and Spearman's ρ) and 95% confidence intervals across length buckets to assess the reliability of model differences.

Reward Models. We utilize two reward models based on xCOMET. The MQM-derived reward signal, referred to as xCOMET-MQM, is generated from error span predictions that identify and classify translation errors by severity. Token-level reward signals are directly computed from these error spans, while sentence-level rewards are obtained as a weighted average of the token-level severity spans. We also use the standard sentence-level reward signal provided by xCOMET as a baseline for comparison.

Training Configurations. We finetune NLLB, Tower and Gemma models using MLE and RL methods detailed in Section 3 with the following configurations:

- **SFT:** A baseline model supervised finetuned on the parallel data using MLE (1).
- **sRL:** We compare using sentence-level xCOMET with BLEU. The learning algorithm used is PPO (3), a current state-of-the-art alignment method for MT.
- **tRL:** We use token-level xCOMET and compare it with partial BLEU, which is based on reward shaping as detailed in Section 4. The learning algorithm used is tPPO (5), the proposed token-level version of PPO.
- **CPO** (Xu et al., 2024b): A state-of-the-art preference optimization learning method for MT, offering a more efficient variant of DPO (Rafailov et al., 2023). We construct the preference dataset by generating multiple outputs from the MT model using the training datasets,[1] and then induce preferences using the xCOMET metric, comparing these outputs to human-written references.

Translate the following text from {**source_lang**} into {**target_lang**}. {**source_lang**}: {**source_sentence**}. {**target_lang**}:

Table 1: Prompt used for Tower and Gemma.

Hyperparameter Details. We use the same hyperparameter settings for NLLB, Tower, and Gemma. We use HuggingFace's Transformers library (Wolf et al., 2020) and the Transformers Reinforcement Learning (TRL) library to facilitate RL training. We perform MLE training with Adam (Kingma and Ba, 2015) as the optimization algorithm, learning rate decay starting from 1×10^{-5} and early stopping. We use PPO with a learning rate of 1.41×10^{-6}, γ set as 0.99, trajectory limit set as 10,000. Mini-batch updates are performed with a batch size of 16 over 4 PPO epochs. The translation prompt for LLM-based MT systems is shown in Table 1.

5.2 Results and Main Findings

We present the main results of comparing the different methods across datasets trained using NLLB in Table 2 and across models in Table 3.

tRL Consistently Outperforms SFT and sRL Methods across Neural Metrics. For all translation directions reported in Table 2, ''tRL w/ xCOMET-MQM'' outperforms SFT and its sentence-level counterpart ''sRL w/ xCOMET-MQM'' across all neural metrics considered. SFT significantly improves translation quality by tailoring the pre-trained MT model to the specific target language pairs. Moreover, applying RL methods (sRL or tRL) on top of SFT further enhances the MT model's performance by directly optimizing translations based on targeted reward signals. When comparing the sRL and tRL methods, we observe that sRL leads to moderate improvements over SFT, while the gains obtained by tRL are more substantial, particularly when assessed with advanced neural metrics. Although

[1] We generate 16 samples with the value of top_p top_k set to 0.9 and 50, respectively.

Model	Metrics					
	BLEU	ChrF	COMET22	xCOMET	BLEURT	CometKiwi-23
			WMT18 EN→DE			
NLLB	41.60 7	66.34 4	86.94 6	94.69 4	76.42 4	72.46 3
+ SFT	43.07 2	67.38 2	87.18 3	94.97 4	76.35 4	72.57 3
+ sRL w/ BLEU	**43.31** 1	**67.57** 1	87.31 2	95.13 3	76.61 3	72.73 2
+ sRL w/ xCOMET	43.09 2	67.34 2	87.18 3	95.55 2	76.72 2	72.84 2
+ sRL w/ xCOMET-MQM	43.03 3	67.29 3	87.10 4	95.50 2	76.72 2	72.40 3
+ tRL w/ BLEU	42.90 4	67.32 3	87.15 3	94.94 4	76.37 4	72.57 3
+ tRL w/ xCOMET-MQM	42.63 5	67.47 2	**88.27** 1	**96.18** 1	**77.60** 1	**75.16** 1
+ CPO w/ xCOMET	42.30 6	66.10 4	87.02 5	95.31 3	76.70 2	73.08 2
			WMT18 DE→EN			
NLLB	43.50 5	66.12 5	86.45 5	93.95 4	75.84 4	73.16 5
+ SFT	45.57 2	67.83 2	87.22 3	95.08 3	76.79 3	74.11 3
+ sRL w/ BLEU	**45.78** 1	**67.97** 1	87.26 2	95.07 3	76.85 2	74.05 3
+ sRL w/ xCOMET	45.51 3	67.78 3	87.21 3	95.28 2	76.76 3	74.06 3
+ sRL w/ xCOMET-MQM	45.55 3	67.80 3	87.21 3	95.17 3	76.80 3	74.11 3
+ tRL w/ BLEU	45.53 3	67.74 3	87.29 2	95.14 3	76.84 2	74.00 4
+ tRL w/ xCOMET-MQM	45.71 2	67.60 4	**87.91** 1	**96.02** 1	**77.70** 1	**74.42** 1
+ CPO w/ xCOMET	43.88 4	66.05 5	87.10 4	95.33 2	77.08 2	74.23 2
			IWSLT2017 EN→FR			
NLLB	43.61 4	65.76 3	84.93 3	90.27 3	72.61 4	70.06 4
+ SFT	45.63 2	67.59 2	86.01 2	91.07 2	74.41 2	71.78 2
+ sRL w/ BLEU	45.77 2	**67.68** 1	86.04 2	91.11 2	74.44 2	71.79 2
+ sRL w/ xCOMET	45.68 2	67.61 2	86.03 2	91.67 2	74.45 2	71.77 2
+ sRL w/ xCOMET-MQM	45.76 2	67.60 2	86.08 2	91.68 2	74.40 2	71.77 2
+ tRL w/ BLEU	45.71 2	67.63 2	86.00 2	91.14 2	74.48 2	71.82 2
+ tRL w/ xCOMET-MQM	**46.58** 1	60.07 4	**87.17** 1	**92.17** 1	**75.58** 1	**72.80** 1
+ CPO w/ xCOMET	43.96 3	65.24 3	85.55 3	91.65 2	74.02 3	71.86 3
			IWSLT2017 FR→EN			
NLLB	45.76 3	65.76 3	87.15 3	94.69 3	77.55 3	71.99 3
+ SFT	48.65 2	67.69 2	88.22 2	95.34 2	78.74 2	73.59 2
+ sRL w/ BLEU	48.76 2	67.76 2	88.22 2	95.33 2	78.74 2	73.54 2
+ sRL w/ xCOMET	48.62 2	**67.88** 1	88.30 2	**95.53** 1	78.73 2	73.56 2
+ sRL w/ xCOMET-MQM	48.61 2	67.66 2	88.31 2	**95.53** 1	78.74 2	73.56 2
+ tRL w/ BLEU	48.56 2	67.69 2	88.21 2	95.33 2	78.73 2	73.60 2
+ tRL w/ xCOMET-MQM	**49.06** 1	61.17 4	**88.46** 1	95.27 2	**79.10** 1	**74.74** 1
+ CPO w/ xCOMET	48.46 2	67.54 2	88.20 2	95.42 2	78.71 2	73.52 2

Table 2: Evaluation of NLLB models on WMT18 (EN↔DE) and IWSLT2017 (EN↔FR), with rows grouped by test set. We provide automatic evaluation metrics for the best base model, baseline (finetuned base model) in each dataset and the variations with sentence-level and token-level RL training. BLEU and xCOMET serve as reward models in the context of RL training. MQM scores are predicted from the error spans ($y = y_{\text{MQM}}$) (Guerreiro et al., 2024). Best-performing values are **bolded**, and models are grouped into statistically significant quality clusters.

Model	Metrics					
	BLEU	ChrF	COMET22	xCOMET	BLEURT	CometKiwi-23
	WMT24 EN→DE					
NLLB	35.01 6	53.92 5	64.30 4	83.40 2	60.90 3	55.50 6
+ SFT	37.92 3	59.29 3	65.20 2	82.50 5	61.41 2	58.20 3
+ sRL w/ BLEU	38.54 2	59.77 2	65.50 2	82.50 5	61.40 2	58.30 3
+ sRL w/ xCOMET	37.88 3	59.16 3	65.50 2	82.80 4	60.40 4	57.90 4
+ sRL w/ xCOMET-MQM	37.14 4	59.29 3	65.40 2	83.30 2	60.40 4	57.50 5
+ tRL w/ BLEU	38.53 2	59.60 2	65.49 2	82.80 4	61.41 2	59.10 2
+ tRL w/ xCOMET-MQM	**39.42** 1	**60.40** 1	**66.43** 1	**83.60** 1	**62.29** 1	**60.70** 1
+ CPO w/ xCOMET	36.07 5	55.32 4	64.70 3	82.90 4	60.80 3	58.10 3
Tower	42.77 6	62.60 5	71.80 5	88.50 4	68.59 5	67.20 4
+ SFT	45.14 4	63.30 4	72.18 4	88.90 3	69.18 4	67.30 4
+ sRL w/ BLEU	46.07 2	63.77 3	72.26 4	89.20 3	70.01 3	68.20 3
+ sRL w/ xCOMET	45.59 3	64.11 2	72.73 2	90.50 2	70.58 2	69.90 2
+ sRL w/ xCOMET-MQM	45.58 3	63.76 3	72.41 3	90.20 2	70.55 2	69.70 2
+ tRL w/ BLEU	46.13 2	64.15 2	72.22 4	89.22 3	70.11 3	68.33 3
+ tRL w/ xCOMET-MQM	**46.92** 1	**65.63** 1	**74.66** 1	**91.90** 1	**71.80** 1	**71.20** 1
+ CPO w/ xCOMET	43.15 5	61.13 6	70.65 6	87.80 5	68.23 5	67.38 4
Gemma	15.13 7	51.99 5	54.84 7	61.40 5	51.35 6	42.80 6
+ SFT	35.19 5	59.03 3	69.86 5	86.10 4	66.13 5	64.30 5
+ sRL w/ BLEU	36.34 3	59.33 2	69.96 5	86.10 4	66.50 4	64.40 5
+ sRL w/ xCOMET	36.16 3	59.28 2	70.82 2	87.20 2	66.90 3	66.30 2
+ sRL w/ xCOMET-MQM	35.98 4	59.33 2	70.39 3	87.10 2	67.20 2	65.70 3
+ tRL w/ BLEU	36.63 2	59.52 1	70.02 4	86.40 3	66.70 3	64.50 5
+ tRL w/ xCOMET-MQM	**36.90** 1	**59.58** 1	**71.12** 1	**88.10** 1	**68.00** 1	**66.90** 1
+ CPO w/ xCOMET	35.05 6	58.52 4	69.22 6	86.40 3	66.80 3	65.10 4

Table 3: Evaluation metrics for NLLB, Tower, Gemma and its variations across WMT24 EN→DE. Best-performing values are **bolded**, and models are grouped into statistically significant quality clusters.

the chrF scores for tRL models trained with xCOMET-MQM are lower on IWSLT2017, this is likely due to a mismatch between the reward signal and the evaluation metric: Token-level rewards optimize semantic quality, not character-level overlap. This can lead to fluent, accurate translations that diverge lexically from references, reducing chrF despite improved overall quality. Neural metrics, which are more robust to surface-level variation and better aligned with human judgments (Freitag et al., 2022, 2023), as well as human evaluation (see Section 5.4) (the gold standard for assessing translation quality) consistently show improvements with our tRL approach. Appendix C presents a focused quantitative and qualitative analysis of cases in which chrF decreases while xCOMET improves.

tRL Improves Translation Quality for LLM-Based MT Systems, Tower and Gemma. Our severity-based, fine-grained mechanism significantly improves translation quality across all automatic evaluation metrics, as shown in Table 3. These findings highlight that tRL not only improves the quality of state-of-the-art MT models but can also significantly boost stronger LLM-based MT systems, demonstrating its broad applicability and potential for advancing multilingual MT systems.

On-Policy PPO Results in Better Translation Quality than RL-Free Method, CPO. Both sentence-level and token-level RL methods achieve higher evaluation scores than CPO, demonstrating significant improvement in translation quality

across language pairs. Unlike CPO, which focuses on maintaining predefined constraints imposed by the preference dataset, RL methods like PPO can flexibly and dynamically adjust the MT model based on real-time feedback from the reward models via iterative feedback and refinement. Furthermore, tRL uses fine-grained reward signals from xCOMET-MQM capturing a wider range of linguistic features and quality indicators, thus offering more precise and contextually relevant feedback during the training process. This feedback can be leveraged more effectively with PPO than with CPO.

xCOMET is a Superior Reward Model than Lexical MT Metrics. The role of the reward model in achieving alignment is crucial, as also evidenced by our findings (Tables 2 and 3). Our results clearly show that using xCOMET as a reward model, particularly at the token level, significantly improves translation quality as measured by several metrics. Given that xCOMET exhibits a strong correlation with human judgments, it proves to be an essential tool for guiding MT models toward higher translation quality. In contrast, traditional metrics like BLEU, based on N-gram overlap, can fall short in aligning with human judgments as they do not capture contextual nuances and semantic understanding (Freitag et al., 2022). Consequently, BLEU performs less effectively compared to neural metrics like xCOMET in this setup which use contextual embeddings. Therefore, incorporating neural metrics as reward models is crucial for capturing the subtleties of language and improving the overall quality and reliability of MT models.

5.3 Ablation Study

We present several ablations to study how the design choices employed impact the learning and the final translation quality of the optimized model.

Choice of Severity Map Impacts Learning. We investigate the impact of different severity maps on token-level RL training using xCOMET, as detailed in Table 4. The severity maps we evaluate include the default MQM-based map (MQM), our custom map (OUR), the reversed MQM-based map (RMQM), the reversed custom map (ROUR), and a binary map (BIN). Our findings, shown in Table 5, highlight the importance of having gradual transitions between reward values. Smooth transition severity map to result in better translation quality. In contrast, abrupt changes in the reward signal can destabilize learning, leading to inconsistent training, oscillations, or convergence to suboptimal policies (Ranzato et al., 2016; Sutton and Barto, 2018). Additionally, we find that the binary severity map, which ignores the severity of errors, provides less informative feedback to the model, resulting in slightly lower performance than maps that offer more nuanced assessments. Although designing custom severity maps can increase complexity and require hyperparameter tuning, our experiments suggest they can be set in a straightforward way with minimal overhead. We leave a more systematic investigation of these mappings, including the possibility of learning them during training, to future work.

WORD	BIN	MQM	RMQM	OUR	ROUR
CORRECT	1	0	25	8	−1
MINOR	−1	−1	5	4	−2
MAJOR	−1	−5	1	2	−4
CRITICAL	−1	−25	0	1	−8

Table 4: Severity maps.

tRL Improves Training Stability over sRL. Figure 3 shows the evolution of mean rewards during training for sRL and tRL across the two datasets for the NLLB system. As observed, tRL training exhibits a more stable and consistently increasing reward trajectory, which is crucial for ensuring steady improvements and reducing the risk of performance-degrading fluctuations or overfitting.

tRL Improves Translation Quality for Longer Sequences. Building on our hypothesis that tRL is particularly effective for longer sequences, we present COMET22 scores for the WMT24 EN→DE dataset in Figure 4, grouped by source sequence length and comparing different training methods, including NLLB and TOWER. The figure also shows the distribution of source sequence lengths in the training and test data. Notably, the training data is skewed toward shorter inputs—a common characteristic of large MT corpora—whereas the WMT24 test set includes a broader distribution with a higher proportion of long sequences. This discrepancy highlights the need for models that generalize well to longer inputs. In this context, ‘‘longer sequences’’ refers

Model	Severity Map	Metrics					
		BLEU	ChrF	COMET22	xCOMET	BLEURT	CometKiwi-23
		IWSLT2017 EN→FR					
NLLB	–	43.61	65.76	84.93	90.27	72.61	70.06
+ SFT	–	45.63	**67.59**	86.01	91.07	74.41	71.78
+ tRL w/ xCOMET-MQM	Bin	45.44	60.43	85.87	90.90	74.23	71.58
+ tRL w/ xCOMET-MQM	MQM	45.68	60.67	85.97	91.01	74.33	71.74
+ tRL w/ xCOMET-MQM	rMQM	41.47	59.74	82.20	84.58	70.45	67.29
+ tRL w/ xCOMET-MQM	Our	**46.58**	60.07	**87.17**	**92.17**	**75.58**	**72.80**
+ tRL w/ xCOMET-MQM	rOur	45.89	60.10	85.90	91.33	74.39	71.79

Table 5: Automatic evaluation metrics for several severity maps setup in the context of token-level RL training. Best-performing values are **bolded**.

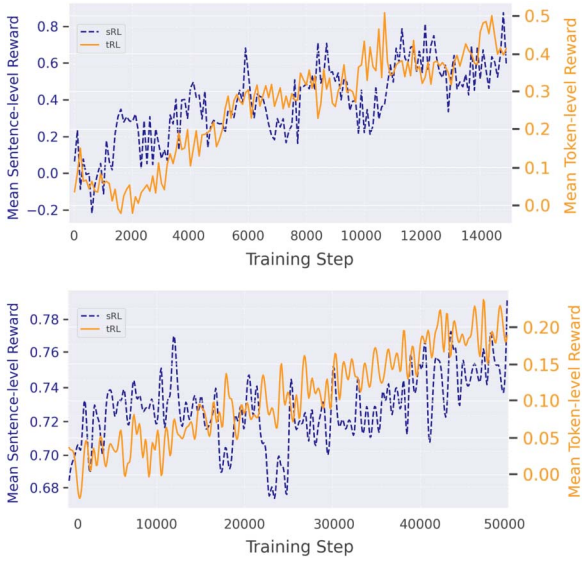

Figure 3: Mean rewards per training step for the IWSLT2017 EN→FR (top) and WMT18 EN→DE (bottom) datasets using xCOMET as the reward model with NLLB. The learning curves highlight training stability trends, where tRL (orange) displays greater stability than sRL (blue). Note that reward scales are not directly comparable due to differences in granularity and clipping methods.

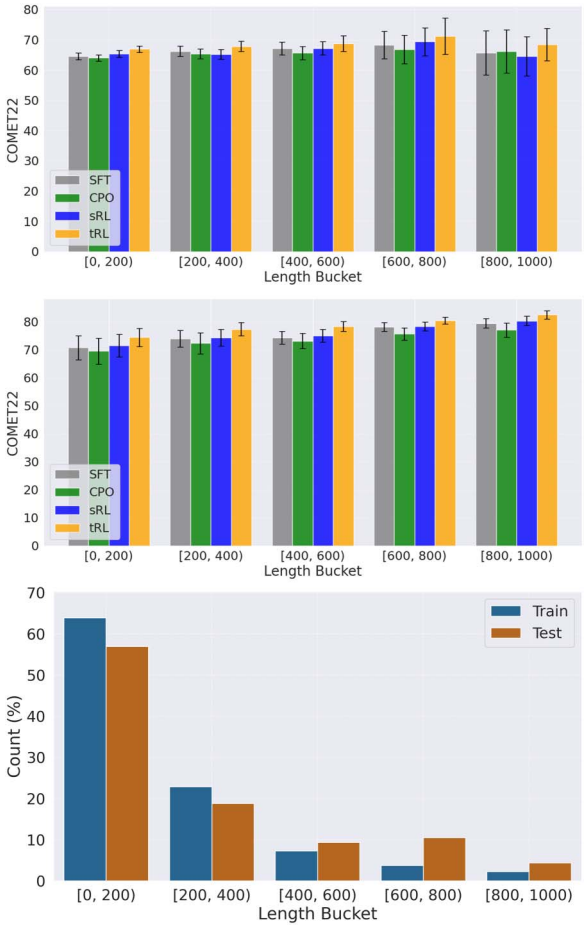

Figure 4: COMET22 scores for NLLB (top), Tower (middle), and a comparative analysis of training and test data length distribution (bottom) on WMT24 EN→DE across increasing source sentence lengths, measured by character string length.

to short paragraphs rather than single sentences. These are not document-level inputs and remain within the 512-token limit of xCOMET, ensuring that the reward model processes them without truncation during training. tRL consistently outperforms other training methods, especially on longer sequences. We attribute this to its ability to capture localized, fine-grained reward signals during training. These findings further support our earlier results: tRL demonstrates the most robust performance, with smaller performance drops as the source sentence length increases, confirming its strength in handling complex sentence structures. For completeness, we also evaluate in Appendix B a hybrid model that applies sRL to short inputs and tRL to long ones. This approach outperforms sRL but remains inferior to tRL.

Model	Metrics					
	BLEU	ChrF	COMET22	xCOMET	BLEURT	CometKiwi-23
	IWSLT2017 EN→FR					
NLLB	43.61	65.76	84.93	90.27	72.61	70.06
+ SFT	45.63	67.59	86.01	91.07	74.41	71.78
+ sREINFORCE w/ xCOMET	45.70	67.58	86.11	91.29	74.43	71.75
+ sREINFORCE w/ xCOMET-MQM	45.72	67.59	86.07	91.35	74.78	71.80
+ tREINFORCE w/ xCOMET-MQM	46.07	60.60	86.24	91.77	74.98	71.87
+ sPPO w/ xCOMET	45.68	**67.61**	86.03	91.67	74.45	71.77
+ sPPO w/ xCOMET-MQM	45.76	67.60	86.08	91.68	74.40	71.77
+ tPPO w/ xCOMET-MQM	**46.58**	60.07	**87.17**	**92.17**	**75.58**	**72.80**

Table 6: Automatic evaluation metrics for REINFORCE and PPO in the context of token-level RL training. Best-performing values are **bolded**.

REINFORCE and PPO Are Suitable Methods for Training MT Systems. We compare REINFORCE and PPO for RL-based MT with xCOMET as the reward model, evaluating their impact on translation quality (Table 6). Both methods are effective, but PPO achieves superior overall metric scores due to features such as objective clipping and KL divergence control, which enhance training stability. However, REINFORCE remains a strong alternative for simpler implementations that aim to achieve competitive performance.

Efficiency and Quality Tradeoffs in Token-Level Reward Computation. Using xCOMET as a reward model for tRL yields higher-quality translations than BLEU, but it also incurs increased computational costs, as detailed in Table 7. Due to its larger pre-trained encoder, xCOMET exhibits significantly higher latency (average seconds per token) and lower throughput (tokens per second). It is worth noting that throughput values can appear higher than latency alone might suggest, as this metric benefits from amortized per-call overheads and batching. The quality improvement, measured by COMET22, reflects the performance of our best model (Tower) on WMT24 EN→DE after fine-grained optimization with each reward model. Despite the slower processing, xCOMET's superior reward quality is particularly valuable in token-level feedback scenarios where high-quality rewards are essential and the additional computational cost is manageable with GPU acceleration. Furthermore, advances in computational efficiency for transformer architectures—such as quantization (Jacob et al., 2018), FlashAttention (Dao et al., 2022; Dao, 2024), and distillation (Hinton et al., 2015)—can help mitigate this computational load, making xCOMET more practical for broader applications.

	Latency ↓	Throughput ↑	Quality ↑
BLEU	2.08×10^{-4}	2.24×10^{5}	72.22
xCOMET	8.17×10^{-2}	7.64×10^{2}	74.66

Table 7: Comparison of BLEU and xCOMET reward models with respect to computational efficiency (latency, throughput) and final translation quality measured by COMET22.

5.4 Human Evaluation

Setup. For our human evaluation, we used Direct Assessments (Graham et al., 2013, DAs) to score translations on a scale from 0 to 100, following the standard WMT human evaluation methodology. We evaluate 200 randomly chosen instances from the WMT18 EN → DE dataset. Two professional translators, both native speakers of the target language, assess the references and NLLB translations using the following methods: SFT, CPO with xCOMET, tRL with xCOMET, our proposed severity mapping, and sRL with xCOMET.

While WMT18 EN → DE allows for easy comparison between several methods due to shorter sequences, we conducted a second human evaluation on the WMT24 EN → DE dataset to validate our empirical finding that tRL benefits longer input sequences. For this setting, we directly compare sRL with xCOMET and tRL with xCOMET in a pairwise setting using outputs from the Tower

	CPO	SFT	tRL	sRL	Reference
Ann1	59.5	64.5	66.9	66.8	76.0
Ann2	53.2	54.2	56.1	57.6	60.0

Table 8: Human evaluation on WMT18.

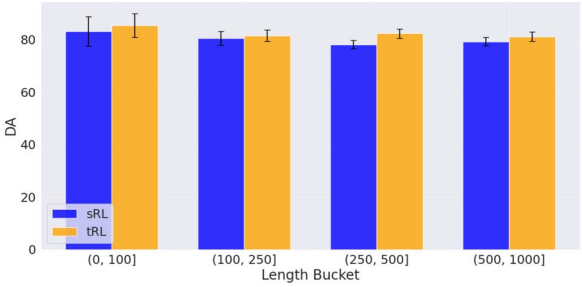

Figure 5: DA scores for sRL and tRL with Tower on WMT24 EN→DE across increasing source sentence lengths.

model. To ensure adequate coverage across different sequence lengths, we performed stratified sampling based on source length, using bins [0, 100, 250, 500, 1000] with 50 instances per bin.

Findings. As shown in Table 8, both sRL and tRL models consistently outperform SFT and CPO, demonstrating their advantage in translation quality. On the WMT24 dataset, tRL achieves an average DA score of 82.6, 2.3 points higher than sRL, with consistent gains across sentence-length buckets, as shown in Figure 5; longer sentences exhibit more significant improvements based on error bar overlap. Human evaluations were conducted by two professional annotators, and inter-annotator agreement is moderate-to-high (Pearson's $r = 0.59$, Spearman's $\rho = 0.57$), indicating reliable scoring. These results align with our automatic evaluation, including the ablation analysis, which collectively shows that tRL enhances stability and translation quality, particularly for longer sentences.

6 Conclusion

In this work, we propose a new method for improving NMT that uses fine-grained reward optimization with xCOMET as a token-level reward model. While exposure bias arises from SFT, sentence-level RL addresses this issue but introduces reward sparsity due to coarse-grained feedback. Our token-level RL approach overcomes this by providing a denser and more informative reward signal to enhance translation quality. Our experiments show that incorporating fine-grained reward mechanisms significantly improves MT quality, especially for longer sequences, and also stabilizes training. Additionally, token-level RL training outperforms sentence-level RL training in most evaluation metrics. Our findings show that fine-grained RL offers a more effective MT optimization framework by mitigating reward sparsity and aligning better with human judgments.

Acknowledgments

We thank the members of SARDINE lab for their useful and constructive comments. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

Sweta Agrawal, José G. C. De Souza, Ricardo Rei, António Farinhas, Gonçalo Faria, Patrick Fernandes, Nuno M. Guerreiro, and Andre Martins. 2024. Modeling user preferences with automatic metrics: Creating a high-quality preference dataset for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14503–14519, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.803

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.662

Duarte Miguel Alves, José Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6401

Meng Cao, Lei Shu, Lei Yu, Yun Zhu, Nevan Wichers, Yinxiao Liu, and Lei Meng. 2024. Enhancing reinforcement learning with dense rewards from language model critic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9138, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.515

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-4012

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clémençon. 2022. What are the best systems? New perspectives on NLP benchmarking. In *Advances in Neural Information Processing Systems*, volume 35, pages 26915–26932. Curran Associates, Inc. https://doi.org/10.52202/068431-1951

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Re. 2022. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*. https://doi.org/10.52202/068431-1189

Rati Devidze, Parameswaran Kamalaruban, and Adish Singla. 2022. Exploration-guided reward shaping for reinforcement learning under sparse rewards. In *Advances in Neural Information Processing Systems*, volume 35, pages 5829–5842. Curran Associates, Inc. https://doi.org/10.52202/068431-0422

Miguel Domingo, Álvaro Peris, and Francisco Casacuberta. 2017. Segment-based interactive-predictive machine translation. *Machine Translation*, 31(4):163–185. https://doi.org/10.1007/s10590-017-9213-3

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for

Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.100

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.51

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.2

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal Archive*, 12:23–38.

Prasoon Goyal, Scott Niekum, and Raymond J. Mooney. 2019. Using natural language for reward shaping in reinforcement learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2385–2391. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/331

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995. https://doi.org/10.1162/tacl_a_00683

Geyang Guo, Ranchi Zhao, Tianyi Tang, Xin Zhao, and Ji-Rong Wen. 2024. Beyond imitation: Leveraging fine-grained quality signals for alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.451

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv e-prints*, page arXiv:1503.02531.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2018.00286

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2024. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared

task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.63

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/D13-1176

Samuel Kiegeland and Julia Kreutzer. 2021. Revisiting the weaknesses of reinforcement learning for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1673–1681, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.133

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.1

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 135–144, Lisboa, Portugal. European Association for Machine Translation.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018a. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-3012

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018b. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1165

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2012

Wendi Li, Wei Wei, Kaihe Xu, Wenfeng Xie, Dangyang Chen, and Yu Cheng. 2024. Reinforcement learning with token-level feedback for controllable text generation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1704–1719, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.111

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring

and describing translation quality metrics. *Tradumàtica: Tecnologies de la traducció*, 0:455–463. https://doi.org/10.5565/rev/tradumatica.77

Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. 2023. Toward human-like evaluation for natural language generation with error analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5892–5907, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.324

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.520

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 278–287, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. 2017. Reinforcement learning for bandit neural machine translation with simulated human feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1474, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1153

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846. https://doi.org/10.1038/s41586-024-07335-x, PubMed: 38839963

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc. https://doi.org/10.52202/068431-2011

Yichen Ouyang, Lu Wang, Fangkai Yang, Pu Zhao, Chenghua Huang, Jianfeng Liu, Bochen Pang, Yaming Yang, Yuefeng Zhan, Hao Sun, Qingwei Lin, Saravan Rajmohan, Weiwei Deng, Dongmei Zhang, Feng Sun, and Qi Zhang. 2024. Token-level Proximal Policy Optimization for Query Generation. *arXiv e-prints*, arXiv:2411.00722. https://doi.org/10.18653/v1/2025.emnlp-main.1589

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.51

Pavel Petrushkov, Shahram Khadivi, and Evgeny Matusov. 2018. Learning from chunk-based feedback in neural machine translation. In

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 326–331, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2052

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Martin L. Puterman. 1990. Chapter 8 Markov decision processes. In *Stochastic Models*, volume 2 of *Handbooks in Operations Research and Management Science*, pages 331–434. Elsevier. https://doi.org/10.1016/S0927-0507(05)80172-0

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc. https://doi.org/10.52202/075280-2338

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André 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. https://doi.org/10.18653/v1/2023.wmt-1.73

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.52, https://doi.org/10.18653/v1/2022.wmt-1.60

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann,

Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. 2016. High-dimensional continuous control using generalized advantage estimation. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*.

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

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. https://doi.org/10.18653/v1/2020.acl-main.704

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1159

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction, 2nd ed.* Adaptive computation and machine learning. The MIT Press, Cambridge, MA, US.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256. https://doi.org/10.1023/A:1022672621406

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1137

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1397

Lijun Wu, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2017. Sequence prediction with unlabeled data by reward function learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3098–3104. https://doi.org/10.24963/ijcai.2017/432

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv e-prints*, arXiv:1609.08144.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, volume 36, pages 59008–59033. Curran Associates, Inc. https://doi.org/10.52202/075280-2574

Han Xia, Songyang Gao, Qiming Ge, Zhiheng Xi, Qi Zhang, and Xuanjing Huang. 2024. Inverse-Q*: Token level reinforcement learning for aligning large language models without preference data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8178–8188, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.478

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *ICLR*.

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

Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Gunsoo Han, Daniel Nam, Daejin Jo, Kyoung-Woon On, Mark Hasegawa-Johnson, Sungwoong Kim, and Chang Yoo. 2024. TLCR: Token-level continuous reward for fine-grained reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14969–14981, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.1080/1358314X.2024.2437866

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024. A preference-driven paradigm for enhanced translation with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3385–3403, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.186

A Details of the Severity Assignment Algorithm

In this section, we provide a detailed description of the severity assignment algorithm used in our approach, focusing on the case of tokens overlapping annotated error spans. Our implementation resolves multiple overlapping error spans by assigning the token the worst severity among all overlapping spans. This choice aligns with MQM annotation practices, where the most severe issue in a region governs its quality classification.

Formally, if a token overlaps spans labeled minor, major, and critical, the token is assigned critical. We avoid averaging or length-weighted schemes to remain fully tokenizer-agnostic. A token is considered affected by a span if it overlaps with it in any way, not only if it is fully contained. This avoids mismatches in cases where tokens are longer than the annotated spans.

Formally, let

- $T = \{t_1, \ldots, t_n\}$ be the set of tokens,
- $S = \{s_1, \ldots, s_m\}$ be the set of annotated error spans,
- $\sigma(s) \in \{\text{minor} < \text{major} < \text{critical}\}$ denote the severity of span s.

The severity assigned to token t_i is defined as:

$$\sigma(t_i) = \begin{cases} \max\{\sigma(s) \mid s \in S,\, t_i \cap s \neq \varnothing\}, \\ \qquad\qquad \text{if } \exists s \in S : t_i \cap s \neq \varnothing \\ \text{None}, \quad \text{otherwise.} \end{cases}$$

Example. Suppose the text is tokenized as a single token ‘‘abc’’, and error spans are defined as $[..a]$ (minor), $[b]$ (major), and $[c..]$ (critical). Since the token ‘‘abc’’ overlaps with all three spans, its assigned severity is critical.

B Details of the Hybrid sRL–tRL Experiment

To examine whether the improvements of tRL are primarily driven by long-sequence behavior, we evaluate a hybrid reinforcement learning (hRL)

Method	Metrics					
	BLEU	ChrF	COMET22	xCOMET	BLEURT	CometKiwi-23
Tower	42.77	62.60	71.80	88.50	68.59	67.20
+ SFT	45.14	63.30	72.18	88.90	69.18	67.30
+ sRL w/ xCOMET-MQM	45.58	63.76	72.41	90.20	70.55	69.70
+ tRL w/ xCOMET-MQM	**46.92**	**65.63**	**74.66**	**91.90**	**71.80**	**71.20**
+ hRL w/ xCOMET-MQM	45.96	65.07	73.20	90.57	71.30	69.34

Table 9: Hybrid RL (hRL) results compared to baselines on WMT24 EN→DE.

approach that applies sentence-level RL (sRL) to short inputs and token-level RL (tRL) to long ones. Short sentences are defined as those below the average source length in the training data, and long sentences as those above it. This experiment follows our best-performing configuration: the Tower model with xCOMET-MQM used as the reward signal. Table 9 reports the results alongside the relevant baselines from Table 3. The hRL model improves over sRL on most metrics but consistently falls short of tRL, indicating that tRL provides the strongest training signal across sentence lengths without the added complexity of a hybrid setup.

C Analysis of chrF Drop Cases

This section provides a focused quantitative and qualitative analysis of translations where chrF decreases but xCOMET improves under token-level RL (tRL). The goal is to examine whether the observed chrF drop corresponds to genuine translation degradation or reflects a metric mismatch. We leverage available human evaluation data for WMT24 EN→DE, using Direct Assessment (DA) scores as reliable indicators of translation quality.

This analysis stems from the observation that, in some settings, tRL outputs show a noticeable drop in chrF (up to 7 points on IWSLT2017) while simultaneously improving xCOMET and other quality metrics. This raised concerns that such a drop might reflect lexical imprecision or other undesirable artifacts. To investigate, we conducted a detailed error analysis focusing on cases with the largest discrepancies between chrF and xCOMET. Examining these high-divergence examples provides the clearest insight into whether chrF drops correspond to real quality issues or simply reflect a mismatch between metrics.

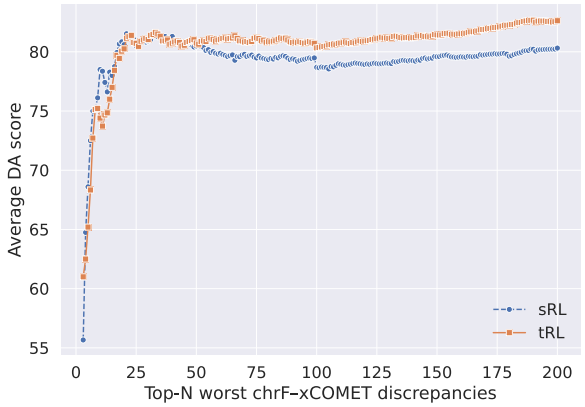

Figure 6: Summary of chrF, xCOMET, and DA for cases with extreme metric discrepancies in WMT24 EN→DE translations.

C.1 Quantitative Analysis

Figure 6 summarizes the extreme discrepancy cases for EN→DE translations. The plot shows the average DA score for the top-N cases with the largest chrF–xCOMET discrepancies. Even with the observed chrF drops and corresponding xCOMET increases, tRL consistently achieves higher human DA scores than sRL, indicating that translation quality is not compromised. This strongly suggests that the apparent chrF decline is a metric artifact rather than a genuine degradation in translation quality.

C.2 Qualitative Analysis

Table 10 presents representative examples of translations with chrF drops but higher xCOMET and DA scores. These cases illustrate how lexical divergence from the reference can lower chrF while yielding translations that are more fluent, semantically accurate, and preferred by human evaluators.

Example 1	
Source	I like both the glasses you posted, these are a really nice colour...makes me want new ones
Reference	Ich mag beide Brillen, die du gepostet hast, die haben eine wirklich nette Farbe ... weckt bei mir den Wunsch, neue zu kaufen
sRL translation	Ich mag beide Brillen, die du gepostet hast, diese sind wirklich eine schöne Farbe... Ich würde gerne neue haben
tRL translation	Mir gefallen beide Brillen, die du gepostet hast, sie haben eine wirklich schöne Farbe... jetzt will ich auch neue
Scores	ChrF 63.4 → 59.7 xCOMET: 96.8 → 99.3 DA: 75.0 → 84.0
Example 2	
Source	I don't have any telemetry for the battery. This is something I might add in a future revision.
Reference	Ich habe keine Telemetrie für den Akku. So etwas könnte ich in Zukunft noch einbauen.
sRL translation	Ich habe keine Telemetrie für den Akku. Das ist etwas, was ich in einer zukünftigen Überarbeitung hinzufügen könnte.
tRL translation	Ich habe keine Telemetrie-Daten für den Akku. Das ist etwas, was ich vielleicht in einer zukünftigen Version hinzufügen könnte.
Scores	ChrF 58.4 → 53.1 xCOMET: 97.2 → 98.2 DA: 84.0 → 94.0
Example 3	
Source	''Which notebook is that?'' Ivory asked, sitting down next to Kari on her bed.
Reference	„Welches Notizbuch ist das?'', fragte Ivory und setzte sich neben Kari auf deren Bett.
sRL translation	„Welches Notizbuch ist das?'' fragte Ivory und setzte sich neben Kari aufs Bett.
tRL translation	„Welches Notizbuch ist das?'', fragte Ivory, während sie sich neben Kari auf das Bett setzte.
Scores	ChrF 85.3 → 80.4 xCOMET: 99.3 → 99.5 DA: 84.0 → 93.0
Example 4	
Source	I'm splurging on a new set of frames, these red ones I reeeeally like.
Reference	Ich verschwende mein Geld für eine neue Brillenfassung, diese rote mag ich seeehr.
sRL translation	Ich habe mir eine neue Brille gekauft, diese rote gefällt mir richtig gut.
tRL translation	Ich gönne mir ein neues Brillengestell, diese roten gefallen mir ganz besonders.
Scores	ChrF 37.8 → 33.8 xCOMET: 95.2 → 95.5 DA: 76.0 → 92.0
Example 5	
Source	''Well my buddys from boot camp here did'' Cohren motioned to Harris, Craith and Ravik. ''But then again, their squad and platoon leader's, they know to wake up before their units.
Reference	„Nun, bei meinem Kumpels aus dem Bootcamp ist das bereits der Fall'', Cohren winkte Harris, Craith und Ravik zu. „Aber auch ihre Truppen- und Zugführer wissen, dass sie vor ihren Einheiten aufwachen müssen.''
sRL translation	„Nun ja, meine Kameraden vom Bootcamp hier haben es getan'', sagte Cohren zu Harris, Craith und Ravik. „Aber ihr Zug- und Platoonführer wissen, dass sie vor ihren Einheiten aufwachen müssen.
tRL translation	„Nun, meine Kameraden vom Bootcamp haben es getan'', sagte Cohren und deutete dabei auf Harris, Craith und Ravik. „Aber sie und ihre Zugführer wissen ja auch, dass sie vor ihren Einheiten aufstehen müssen.''
Scores	ChrF 57.8 → 56.2 xCOMET: 81.9 → 85.1 DA: 74.0 → 85.0

Table 10: Representative translation examples where tRL outputs exhibit lower chrF than sRL but higher xCOMET and human DA scores, illustrating that the chrF drop reflects a metric mismatch rather than an actual decline in translation quality.