# 💫SSR-Zero: Simple Self-Rewarding Reinforcement Learning for Machine Translation

**Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, Sitong Wang**[†]
Tencent Hunyuan, Columbia University[†]
leonzxyang@tencent.com

## Abstract

Large language models (LLMs) have recently demonstrated remarkable capabilities in machine translation (MT). However, most advanced MT-specific LLMs rely heavily on external supervision during training, such as human-annotated reference data or trained reward models (RMs), which are expensive to obtain and difficult to scale. To address this limitation, we propose Simple Self-Rewarding (💫**SSR**), a reinforcement learning (RL) framework for MT that is reference-free and relies solely on self-judging rewards. Using only 13K monolingual examples and Qwen-2.5-7B as the backbone, SSR-Zero-7B outperforms existing MT-specific LLMs as well as larger general LLMs such as Qwen2.5-32B-Instruct on English ↔ Chinese translation benchmarks including WMT23, WMT24, and FLORES200. It further demonstrates strong generalization to low-resource language pairs. In addition, when augmented with external supervision from COMET, our strongest model, SSR-X-Zero-7B, surpasses all existing open-source models under 72B parameters and performs competitively with leading closed-source systems in English ↔ Chinese translation. Our analysis highlights the effectiveness and generalizability of the self-rewarding mechanism relative to external LLM-as-a-judge approaches and demonstrates its complementary benefits when combined with trained RMs. We will publicly release our code, data, and models.

## 1 Introduction

Large language models (LLMs) have recently achieved substantial progress in machine translation (MT) (Aryabumi et al., 2024; Rei et al., 2024b; Cui et al., 2025), benefiting from large-scale pre-training and effective transfer of multilingual knowledge. MT-specific LLMs such as Tower and X-ALMA further improve translation quality through continual pre-training (CPT) on billions of parallel and monolingual tokens, followed by fine-tuning on high-quality human-annotated data (Alves et al., 2024; Cui et al., 2025). While effective, this paradigm relies heavily on parallel data, which – even when available at scale – often suffers from noise and semantic misalignment (Meng et al., 2024), machine-generated contamination (Thompson et al., 2024), and translationese artifacts (Koppel and Ordan, 2011), limiting its reliability as a supervision signal (Uhlig et al., 2025).

In parallel, recent advances in inference-time reasoning have shown that reinforcement learning (RL) can substantially enhance LLM capabilities. Models such as OpenAI o1 (Jaech et al., 2024) and DeepSeek R1 (Guo et al., 2025) employ R1-style training with RL algorithms (e.g., GRPO (Shao et al., 2024), DAPO (Yu et al., 2025)) to incentivize reasoning behaviors, achieving strong performance in tasks such as logic, coding, and mathematics (Guo et al., 2025; Xie et al., 2025; Song et al., 2025). Recent work has begun extending these ideas to MT, either by introducing explicit reasoning patterns (Wang et al., 2024, 2025) or by allowing models to learn reasoning implicitly during training (Feng et al., 2025). However, existing RL-based MT approaches still depend heavily on external supervision, either in the form of human-annotated references or trained reward models distilled from expensive labeled data, which limits their scalability.

To address this limitation, we propose Simple Self-Rewarding (**SSR**), a RL framework for MT that eliminates the need for any external supervision. SSR adopts a self-judging mechanism in which the LLM itself evaluates its translation outputs and produces reward signals, which are then used to optimize the model via GRPO. Using only 13K monolingual sentences (6.5K English and 6.5K Chinese), we train uninstructed Qwen2.5-7B and 3B models, resulting in SSR-Zero-7B and SSR-Zero-3B. SSR-Zero-7B improves its backbone by 18.11% on Chinese-to-English and 14.74%



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 6039–6052

on English-to-Chinese translation.

Extensive experiments on WMT23, WMT24, and FLORES-200 show that SSR-Zero-7B outperforms existing MT-specific LLMs such as TowerInstruct-13B and GemmaX-28-9B, as well as larger general-purpose LLMs including Qwen-2.5-32B-Instruct. When augmented with external COMET rewards, our strongest model, SSR-X-Zero-7B, achieves the best performance among evaluated open-source LLMs under 72B parameters for English $\leftrightarrow$ Chinese translation and performs competitively with closed-source systems such as GPT-4o and Gemini 1.5 Pro. Results on SSR-Zero-3B further indicate that SSR generalizes to smaller base models with weaker judging capabilities.

Finally, we conduct detailed analyses to examine the effectiveness and generalizability of self-rewarding. These include experiments on low-resource languages (Gujarati and Kazakh), comparisons between self-rewarding and external reward models, and an investigation into the impact of reference-based versus referenceless rewards.

In summary, **our contributions are**: 1) We propose SSR, a self-rewarding RL framework for MT that removes reliance on external reward models and reference translations. 2) We demonstrate that SSR substantially improves MT quality across model sizes, language pairs, and resource settings, outperforming strong open-source MT baselines. 3) We show that self-generated rewards complement external rewards, enabling SSR-X-Zero-7B to achieve the strongest performance among evaluated open-source models for English $\leftrightarrow$ Chinese translation. 4) We provide a systematic analysis of reward design choices for RL-based MT and release our code, data, and models to support future research.

## 2 Related Work

### 2.1 Machine Translation with LLMs

Recent advances in large language models (LLMs) have substantially improved machine translation (MT) across many language pairs (Costa-Jussà et al., 2022; Lu et al., 2024; Workshop et al., 2022). Many strong MT-focused LLMs (Rei et al., 2024b; Cui et al., 2025) rely on continual pre-training (CPT) over large-scale mixtures of parallel and monolingual data, often exceeding tens of billions of tokens.

Beyond data scale, prior work has shown that enriching training objectives can further enhance MT performance. For example, Rei et al. (2024a) incorporate auxiliary tasks such as translation evaluation, MQM-style error detection, and named entity recognition, while Cui et al. (2025) propose a sequential data mixing strategy that prioritizes parallel data during CPT, yielding competitive performance with commercial systems such as Google Translate and GPT-4-turbo.

Despite their strong performance, these approaches depend heavily on large volumes of curated or annotated data. As training scales, the cost and availability of such resources increasingly limit the sustainability of MT model development.

### 2.2 MT via Reinforcement Learning

Reinforcement learning (RL) has long been explored in MT to mitigate exposure bias in supervised training (Bengio et al., 2015). Early work applied algorithms such as REINFORCE (Ranzato et al., 2015), actor–critic methods (Bahdanau et al., 2016), and policy gradients (Yu et al., 2017), using either rule-based metrics (e.g., BLEU, ROUGE) or trained reward models (Wu et al., 2017).

More recently, in the context of LLMs, R1- and R1-Zero-style training has demonstrated that RL algorithms such as GRPO, combined with verifiable rewards, can substantially improve reasoning ability (Guo et al., 2025). This paradigm has been extended to MT, either by introducing explicit reasoning patterns or by allowing models to learn latent reasoning processes during RL training.

For example, He et al. (2025) train MT models using manually designed chain-of-thought data with COMET-based rewards, while Feng et al. (2025) explore BLEU-, COMETKiwi-, and hybrid reward signals, achieving strong results with MT-R1-Zero-Sem. Wang et al. (2025) further propose using a large LLM-based judge to evaluate both reasoning steps and translation outputs during RL, targeting literary translation.

Nevertheless, existing RL-based MT methods still rely on external supervision signals, such as reference translations, trained reward models, or large frozen judges. Beyond cost, available parallel data often suffers from noise (Meng et al., 2024), machine-generated contamination (Thompson et al., 2024), and translationese artifacts (Koppel and Ordan, 2011), limiting its reliability as supervision. Recent work has also explored human-preference alignment for MT via RLHF (Moura Ramos et al., 2024; Xu et al., 2024)

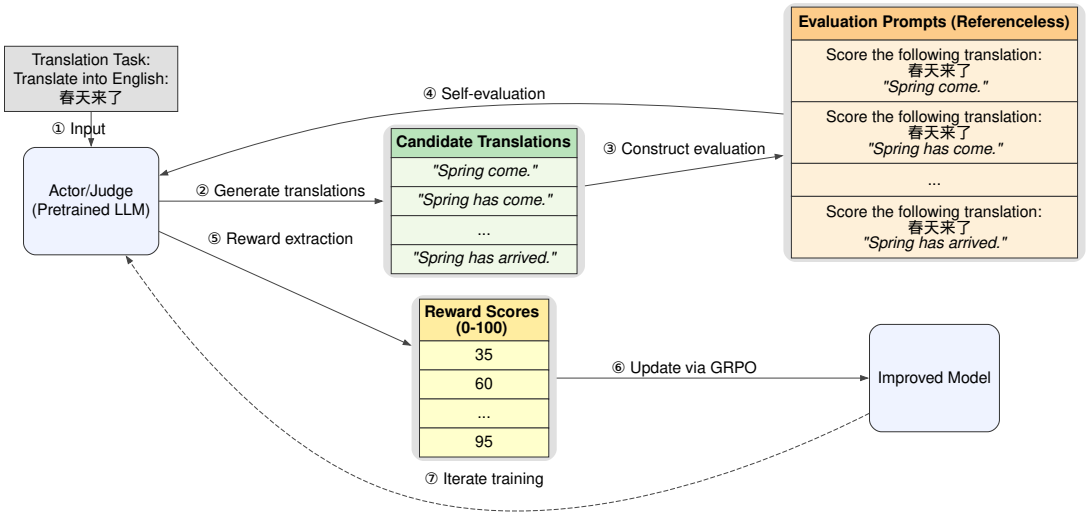


Figure 1: Overview of the 💫SSR framework. SSR is an R1-Zero-like RL training method for machine translation, which uses the same model as both actor and judge. It does not require external reward models or human-annotated reference data. Prompts shown here are simplified for clarity.

and direct quality optimization (Uhlig et al., 2025), but these approaches still depend on annotated preference data.

### 2.3 Self-Judging in RL

Recent work has investigated self-rewarding or self-judging mechanisms, in which LLMs generate their own feedback signals for training (Chen et al., 2024; Wu et al., 2024; Zhang et al., 2025b). Such approaches aim to reduce reliance on human annotations or reward models distilled from human judgments. For instance, Chen et al. (2024) iteratively perform self-instruction sampling, self-judging, and DPO training, demonstrating improvements in both instruction-following and evaluation capabilities.

Related self-improving paradigms, including self-play and self-judging, have shown effectiveness in domains such as mathematical reasoning (Zhang et al., 2025a; Zhao et al., 2025), vision–language alignment (Zhou et al., 2024), and cross-lingual transfer (Chen et al., 2024; Geng et al., 2024; Yang et al., 2024b). However, self-judging remains relatively underexplored for MT.

One notable exception is Zou et al. (2025), who propose a self-play framework based on Monte Carlo Tree Search to derive preferences from cross-lingual semantic consistency. While effective, their approach does not outperform MT-specific LLMs such as TowerInstruct when using the same base model.

In contrast, our approach eliminates the need for external supervision, operates fully online, and achieves strong performance using only monolingual data. These results suggest that sufficiently strong pre-trained LLMs already possess usable translation and MT-evaluation capabilities, pointing toward a viable path for self-improving MT without human feedback.

## 3 Methodology

In this section, we first outline the SSR methodology (§3.1), followed by an introduction of the reward design within the RL framework (§3.2). Finally, we introduce the RL algorithm employed in our work (§3.3).

### 3.1 Simple Self-Rewarding (SSR)

SSR is a R1-Zero-like RL approach with a novel self-evaluation mechanism that simplifies reward signal acquisition. This mechanism leverages a pre-trained LLM that alternates between acting as both an actor and a judge.

As illustrated in Figure 1, the pretained model, at each training step, first plays the role of an actor that accepts a batch of translation prompts (①). For each prompt, the model generates a group of N candidate translations (②). These candidate translations are then constructed on LLM-as-a-judge prompts separately (③). Next, the model switches to a judge role, evaluating all prompts to estimate translation quality and generate judgments (④). Each judgment includes a score from 0 to 100, where 0 indicates poor translation and 100 indi-

cates perfect translation. We extract reward scores from judgments using regular expressions (⑤) and then use them in the RL algorithm (i.e., GRPO) to update the actor model's parameters (⑥). In total, one translation prompt generates N candidate translations and N reward scores. We iterate Step ① through ⑥ multiple times until the model's performance converges (⑦).

Below are the prompts for generating translations (i.e., *actor prompt*) and evaluations (*judge prompt*) used in SSR training. The actor prompt builds on Deepseek-R1-Zero's system prompt (Guo et al., 2025), requiring the model to answer within a specific format (i.e., `<answer></answer>`) and think before responding.

**Actor Prompt: Generating Translations**

A conversation between User and Assistant. The User asks a question, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the User with the answer. The reasoning process is enclosed within `<think> </think>` and answer is enclosed within `<answer> </answer>` tags, respectively, i.e., `<think>` reasoning process here `</think> <answer>` answer here `</answer>`.

User:
Translate the following text to {tgt_lang}:
{src_text}
Assistant:

**Judge Prompt: Self-Evaluating**

A conversation between User and Assistant. The User asks a question, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the User with the answer. The reasoning process is enclosed within `<think> </think>` and answer is enclosed within `<answer> </answer>` tags, respectively, i.e., `<think>` reasoning process here `</think> <answer>` answer here `</answer>`.

User:
Score the following translation from {src_lang} to {tgt_lang} on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".
Additionally, give a score of zero if the translation 1) contains irrelevant content, such as interpretations of the translation, 2) does not match the target language, 3) contains multiple translations.

{src_lang} source: {src_text}
{tgt_lang} translation: {translated_text}
Assistant:

The judge prompt is modified from GEMBA-DA (Kocmi and Federmann, 2023), a widely-used LLM-as-a-judge template for direct assessment of translation, which achieved SOTA performance in translation quality assessment using GPT-4. Compared to GEMBA-DA, our judge prompt includes an "think-before-answer" system instruction. This addition explicitly encourages the model to take advantage of the reasoning capabilities acquired during RL training when evaluating translations. Additionally, we instruct the judge to give a zero score for unwanted candidate translations containing irrelevant content or language misalignment. During training, only the content within `<answer></answer>` tags is extracted and incorporated into the judge's instructions.

### 3.2 Reward Modeling

Our RL training utilizes two types of rewards: *self-reward* and *format reward*.

**Self Reward** This reward estimates the quality of the model's translation using the training model itself, denoted by:

$$r_{\text{self}} = \frac{M_{\text{self}}(\text{src}, \text{trans})}{100}, r_{\text{self}} \in [0, 1]$$

where $M_{\text{self}}$ is the model during the training. Using the judge prompt, the model takes both source text and model translation (without reference translations) and generates a judgment containing a score on a 100-point scale. All rewards are then linearly rescaled to $[0, 1]$ before GRPO.

**Format Reward** This reward checks whether the model generation follows the format defined in the actor prompt:

$$r_{\text{format}} = \begin{cases} 1, & \text{if format is correct} \\ 0, & \text{if format is incorrect} \end{cases}$$

**Overall Reward** In training, we combine the two types of rewards to train our SSR-Zero model:

$$r_{\text{all}} = \begin{cases} r_{\text{self}} + r_{\text{format}}, & \text{if } r_{\text{format}} \neq 0 \\ 0, & \text{if } r_{\text{format}} = 0 \end{cases}$$

In addition, we investigate integrating external reward signals to further enhance model performance. Our strongest model, SSR-X-Zero (SSR with eXternal rewards), incorporates rewards computed by COMET, an automatic MT evaluation metric (Rei et al., 2022) that scores translation quality using source sentences, machine-generated translations, and reference translations:

$$r'_{\text{all}} = \begin{cases} r_{\text{self}} + r_{\text{COMET}} + r_{\text{format}}, & \text{if } r_{\text{format}} \neq 0 \\ 0, & \text{if } r_{\text{format}} = 0 \end{cases}$$

$$r_{\text{COMET}} = M_{\text{COMET}}(\text{src}, \text{trans}, \text{ref}), r_{\text{COMET}} \in [0, 1]$$

### 3.3 RL algorithm

We follow the work of Shao et al. (2024) and Guo et al. (2025) by adopting the Group Related Policy Optimization (GRPO) algorithm for training, as it demonstrates stability and strong performance. Specifically, for each given translation prompt $p$, the policy model $\pi_{\theta_{\text{old}}}$ first samples a group of candidate translations $G$ $\{o^i\}_{i=1}^G$. Then, using the same policy model, we perform the SSR procedure described earlier to obtain rewards $\{r^i_{\text{all}}\}_{i=1}^G$ for all candidate translations. Next, we compute the advantage for the $i$-th candidate translation by normalizing the group-level rewards:

$$A_i = \frac{r^i_{\text{all}} - \text{mean}\left(\{r^i_{\text{all}}\}_{i=1}^G\right)}{\text{std}\left(\{r^i_{\text{all}}\}_{i=1}^G\right)}$$

Using these advantages, GRPO optimizes the policy by maximizing the following objective:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(O|p)} \left[ \frac{1}{G}\sum_{i=1}^{G} \min\left( \frac{\pi_\theta(o^i|p)}{\pi_{\theta_{\text{old}}}(o^i|p)} A_i, \text{clip}\left( \frac{\pi_\theta(o^i|p)}{\pi_{\theta_{\text{old}}}(o^i|p)}, 1-\varepsilon, 1+\varepsilon \right) A_i - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right) \right]$$

where $\varepsilon$ and $\beta$ are hyperparameters, $\pi_{\text{ref}}$ is the reference model, and $D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ is the KL divergence between $\pi_\theta$ and $\pi_{\text{ref}}$.

## 4 Experiments

### 4.1 Experimental Setup

**Dataset** In this paper, we focus on bidirectional translation between English and Chinese, with potential expansion to other language pairs in future work. We use the training dataset released by Feng et al. (2025), originally collected from WMT 2017 through WMT 2020 for English-Chinese sentence pairs. Following their preprocessing, sentences shorter than 30 characters were filtered out. Unlike the original bilingual setup, we use these data monolingually, splitting sentence pairs into separate English and Chinese examples to serve as monolingual source sentences for training. The resulting dataset comprises 13,130 monolingual examples (6,565 in English and 6,565 in Chinese).

For testing, we evaluate the translation performance on the English-to-Chinese (EN-ZH) and Chinese-to-English (ZH-EN) benchmarks of WMT23[1], WMT24[2], and FLORES-200 (Costa-Jussà et al., 2022).

**Metrics** Following the settings in Rei et al. (2024b), we adopt two widely used automatic MT-evaluation metrics: the reference-based XCOMET-XXL metric (Guerreiro et al., 2024), and the reference-free COMETKIWI-XXL metric (Rei et al., 2023), both in their largest available model size.

**Baselines** We compare our models with the following baseline model categories:

**Closed-source models**, including GPT-4o-20241120 (Hurst et al., 2024), Claude-3.5-Sonnet-20240620 (Anthropic, 2024), and Gemini-1.5-Pro.

[1]https://www2.statmt.org/wmt23/translation-task.html
[2]https://www2.statmt.org/wmt24/translation-task.html

| Models | ZH→EN | | | | | | | EN→ZH | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WMT23 | | WMT24 | | Flores200 | | Avg. | WMT23 | | WMT24 | | Flores200 | | Avg. |
| | KIWI | XCM | KIWI | XCM | KIWI | XCM | | KIWI | XCM | KIWI | XCM | KIWI | XCM | |
| **Closed-Source LLMs** | | | | | | | | | | | | | | |
| Claude-3.5-Sonnet | 81.61 | 93.06 | 81.06 | 90.54 | 89.41 | 97.68 | **88.89** | **80.15** | 92.00 | 80.00 | 86.31 | 89.47 | 94.32 | 87.04 |
| GPT-4o | 80.92 | 92.15 | 79.90 | 89.06 | 88.94 | 96.50 | 87.91 | 76.71 | 88.56 | 77.42 | 83.95 | 88.30 | 93.30 | 84.71 |
| Gemini-1.5-Pro | 80.71 | 92.44 | 79.02 | 88.90 | 88.15 | 97.32 | 87.76 | 79.80 | 91.95 | 79.54 | 87.11 | 89.30 | 94.54 | 87.04 |
| **Open-Source LLMs** | | | | | | | | | | | | | | |
| **General Purpose LLMs** | | | | | | | | | | | | | | |
| Qwen3-32B🤔 | 79.74 | 90.79 | 79.20 | 88.47 | 87.68 | 95.75 | 86.94 | 76.94 | 89.75 | 76.96 | 84.10 | 87.45 | 92.18 | 84.56 |
| Qwen3-32B | 80.28 | 91.95 | 79.95 | 89.53 | 88.88 | 97.18 | 87.96 | 79.27 | 91.28 | 79.51 | 86.63 | 89.69 | 94.07 | **86.74** |
| Qwen3-8B🤔 | 78.30 | 89.03 | 77.99 | 86.94 | 85.82 | 93.89 | 85.33 | 74.94 | 88.22 | 75.39 | 82.25 | 86.08 | 91.02 | 82.98 |
| Qwen3-8B | 79.87 | 91.42 | 79.58 | 89.02 | 88.61 | 96.55 | 87.51 | 78.59 | 90.90 | 78.71 | 85.31 | 88.90 | 93.30 | 85.95 |
| Qwen2.5-72B-Instruct | 80.62 | 92.14 | 80.46 | 90.06 | 88.90 | 97.28 | **88.24** | 78.18 | 91.34 | 78.18 | 85.13 | 88.04 | 93.20 | 85.68 |
| Qwen2.5-32B-Instruct | 77.73 | 89.28 | 78.77 | 88.69 | 87.13 | 95.50 | 86.18 | 77.73 | 90.23 | 78.77 | 83.48 | 87.13 | 91.99 | 84.89 |
| Qwen2.5-3B-Instruct | 73.52 | 86.60 | 75.82 | 85.03 | 85.46 | 93.41 | 83.31 | 66.78 | 84.34 | 67.67 | 76.12 | 78.79 | 85.19 | 76.48 |
| Qwen2.5-7B-Instruct | 77.56 | 89.40 | 76.71 | 87.12 | 86.28 | 94.06 | 85.19 | 73.81 | 88.11 | 72.98 | 80.93 | 85.18 | 89.90 | 81.82 |
| QwQ-32B🤔 | 74.61 | 85.12 | 75.08 | 84.34 | 80.88 | 89.21 | 81.54 | 77.33 | 89.10 | 78.13 | 85.03 | 86.51 | 90.93 | 84.51 |
| Gemma2-27B-it | 80.32 | 91.96 | 79.42 | 89.14 | 88.64 | 96.72 | 87.70 | 76.95 | 90.50 | 77.38 | 84.17 | 87.79 | 92.51 | 84.88 |
| Gemma2-9B-it | 79.86 | 91.21 | 79.25 | 88.41 | 88.32 | 96.25 | 87.22 | 75.22 | 89.66 | 74.15 | 81.65 | 85.95 | 90.90 | 82.92 |
| **MT-Specific LLMs** | | | | | | | | | | | | | | |
| TowerInstruct-7B-v0.2 | 77.78 | 89.13 | 76.96 | 85.98 | 86.95 | 94.88 | 85.28 | 73.53 | 87.46 | 70.87 | 77.53 | 84.39 | 88.57 | 80.39 |
| TowerInstruct-13B-v0.1 | 78.53 | 89.90 | 77.57 | 87.12 | 87.30 | 95.80 | 86.04 | 75.56 | 89.28 | 73.81 | 80.81 | 86.22 | 90.69 | 82.73 |
| DeepTrans-7B🤔 | / | / | / | / | / | / | / | 80.01 | 89.00 | 78.89 | 83.85 | 89.23 | 92.85 | 85.64 |
| GemmaX2-28-9B-v0.1 | 79.40 | 90.63 | 78.71 | 88.60 | 87.85 | 96.33 | 86.92 | 77.10 | 90.68 | 75.88 | 83.33 | 87.58 | 92.83 | 84.57 |
| **Ours** | | | | | | | | | | | | | | |
| Qwen2.5-3B | 44.23 | 65.94 | 41.66 | 55.16 | 51.80 | 65.94 | 54.12 | 19.81 | 66.64 | 23.99 | 57.69 | 23.31 | 69.16 | 43.43 |
| SSR-Zero-3B | 77.51 | 89.88 | 77.81 | 86.18 | 87.23 | 95.64 | 85.71 | 73.88 | 87.14 | 73.57 | 79.40 | 85.43 | 88.66 | 81.35 |
| Qwen2.5-7B | 62.62 | 75.69 | 69.04 | 77.33 | 73.62 | 85.54 | 73.97 | 68.25 | 81.63 | 64.28 | 69.48 | 82.00 | 86.07 | 75.29 |
| SSR-Zero-7B🤔 | 79.29 | 92.04 | 79.04 | 89.19 | 87.97 | 96.70 | 87.37 | 79.69 | 91.18 | 79.34 | 85.34 | 89.25 | 93.52 | 86.39 |
| SSR-X-Zero-7B🤔 | 80.62 | 91.92 | 80.56 | 89.42 | 88.84 | 96.62 | 88.00 | 81.11 | 91.56 | 79.67 | 86.75 | 90.08 | 93.98 | **87.19** |

Table 1: Translation quality measured by COMETKIWI-XXL (KIWI) and XCOMET-XXL (XCM) in English-Chinese directions (EN ↔ ZH). **Bold and underlined** indicates the best-performing model, **bold only** the second-best, and underlined only the third-best. "🤔" denotes reasoning models or models operating in thinking mode.

**Open-source general-purpose LLMs**, including the Qwen3 series (Yang et al., 2025) (Qwen3-32B, Qwen3-8B), Qwen2.5 series (Yang et al., 2024a) (Qwen2.5-72B-Instruct, Qwen2.5-32B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-7B), Qwen's reasoning model QwQ-32B (Team, 2025), and the Gemma2 series (Team et al., 2024) (Gemma2-27B-it and Gemma2-9B-it).

**Open-source MT-specific LLMs**, including the Tower series (Alves et al., 2024) (TowerInstruct-7B-v0.2 and TowerInstruct-13B-v0.1), GemmaX2-28-9B-v0.1 (Cui et al., 2025), and DeepTrans-7B (Wang et al., 2025).

**Implementation Details** We use Qwen2.5-7B and Qwen2.5-3B as the backbone model and adopt the GRPO algorithm implemented in the verl[3] framework. All experiments share the same training settings: a batch size of 128, constant learning rate of 5e-7, rollout number of 16, sampling temperature of 1.0 for generation, and temperature of zero when judging. We set the maximum generation length to 1024 tokens during training. Both KL and entropy coefficients of GRPO are set to zero, as we observed better performance with this configuration. All models are trained for four epochs using eight GPUs, each providing 148 TFLOPs of computational power when optimizing models with BF16 precision. For training SSR-X-Zero-7B, we add an additional GPU to serve the COMET model. Training SSR-Zero-7B takes about 17 hours, while SSR-X-Zero-7B training takes 42 hours in total.

[3] https://github.com/volcengine/verl

## 4.2 Main Results

Table 1 shows that our SSR-Zero-7B model performs strongly in translation compared to existing open-source models. It achieves an average score of 87.37 in ZH→EN, outperforming all MT-specific baselines and several larger general-purpose LLMs such as Gemma2-9B-it and Qwen2.5-32B-Instruct. In EN→ZH, it scores 86.39, surpassing all open-source baselines except Qwen3-32B.

| Models | EN→xx | | | | | | xx→EN | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EN→GU | | EN→KK | | EN→xx | | GU→EN | | KK→EN | | xx→EN | |
| | KIWI | XCM | KIWI | XCM | KIWI | XCM | KIWI | XCM | KIWI | XCM | KIWI | XCM |
| Qwen2.5-7B-Instruct | 19.36 | 23.18 | 18.33 | 10.10 | 18.85 | 16.64 | 66.95 | 50.96 | 68.11 | 30.80 | 67.53 | 40.88 |
| Qwen2.5-7B | 15.01 | 21.78 | 16.26 | 10.25 | 15.64 | 16.02 | 48.02 | 44.01 | 58.50 | 26.18 | 53.26 | 35.10 |
| SSR-Zero-7B | 23.47 | 29.30 | 79.05 | 59.06 | **51.26** | **44.18** | 68.31 | 53.84 | 70.68 | 30.81 | **69.50** | **42.33** |

Table 2: Translation quality across low-resource language pairs, including Gujarati (GU), Kazakh (KK), and aggregated results (xx), measured by COMETKIWI-XXL (KIWI) and XCOMET-XXL (XCM).

Compared to closed-source models, SSR-Zero-7B slightly lags in ZH→EN but outperforms GPT-4o in EN→ZH. It significantly improves upon its backbone model (Qwen2.5-7B) by 18.11% in ZH→EN and 14.74% in EN→ZH.

Our strongest model, SSR-X-Zero-7B, achieves new SOTA performance among open-source models under 72B parameters, with scores of 88.00 in ZH→EN and 87.19 in EN→ZH. It only slightly trails Qwen2.5-72B-Instruct (88.24) in ZH→EN.

Furthermore, SSR-Zero-3B demonstrates the effectiveness of our approach on smaller base models, improving Qwen2.5-3B's performance by over 58% in both translation directions.

These results demonstrate the effectiveness and generalizability of leveraging self-generated rewards and external reward models to enhance MT performance.

## 5 Analysis

Although SSR and its combination with external reward models (RMs) effectively enhance MT performance, two research questions (RQs) remain unclear: 1) *Can SSR generalize to other languages, especially low-resource ones?* 2) *How does self-rewarding compare with widely used external RMs?* 3) *How does the inclusion of reference data in RMs affect the final translation performance?* To clarify these points, we conducted a detailed analysis, presented below.

### 5.1 RQ1: SSR for Low-Resource Languages

To evaluate the generalizability of SSR for low-resource languages, we selected Kazakh (KK) and Gujarati (GU) from the WMT19 dataset. We sampled 3k monolingual sentences per language from the WMT19 training set for training and evaluated the models using the entire test set (2k per language).

**Result** Table 2 shows significant performance gains from SSR training based on the Qwen2.5-7B model. For EN → xx translation, COMETKIWI scores improved substantially from 15.64 to 51.26 (+227%), and XCOMET scores rose from 16.02 to 44.18 (+175%). We note that these large relative gains partly reflect the low absolute baselines typical of low-resource language pairs. For xx→EN translation, COMETKIWI scores increased from 53.26 to 69.50 (+30.49%), while XCOMET scores improved from 35.10 to 42.33 (+20.59%). SSR-Zero-7B significantly outperformed Qwen2.5-7B-Instruct, indicating that our approach effectively generalizes and is particularly beneficial in low-resource scenarios.

### 5.2 RQ2: SSR vs. External Reward Models

Specifically, we compare our method with two categories of external frozen RMs: 1) MT-evaluation trained RMs, including COMET[4] and COMETKIWI[5], and 2) LLM-based judge RMs, including Qwen2.5-7B and Qwen2.5-7B-Instruct, using the same judge prompts employed by SSR.

**Results** The evaluation results are summarized in Table 3. As expected, models trained with specialized MT-evaluation RMs (i.e., COMET or COMETKIWI) outperform SSR-Zero-7B – which relies solely on intrinsic judgments from the training model – in average EN → ZH translation scores. Additionally, these specialized RMs also outperform all methods using external LLM-as-a-judge approaches based on the 7B-sized Qwen2.5 model. This indicates that dedicated RMs trained on large annotated datasets possess stronger MT evaluation capabilities compared to general-purpose LLMs such as Qwen2.5-7B(-Instruct). Nevertheless, the SSR mechanism provides complementary benefits. This is evidenced by SSR-X-Zero-7B, which integrates self-rewarding with COMET supervision, still achieves the highest scores in both translation directions.

[4] https://huggingface.co/Unbabel/wmt22-comet-da

[5] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

| Models | ZH→EN | | | | | | | EN→ZH | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WMT23 | | WMT24 | | Flores200 | | Avg. | WMT23 | | WMT24 | | Flores200 | | Avg. |
| | KIWI | XCM | KIWI | XCM | KIWI | XCM | | KIWI | XCM | KIWI | XCM | KIWI | XCM | |
| Qwen2.5-7B | 62.62 | 75.69 | 69.04 | 77.33 | 73.62 | 85.54 | 73.97 | 68.25 | 81.63 | 64.28 | 69.48 | 82.00 | 86.07 | 75.29 |
| **w/ External trained MT-evaluation RM:** | | | | | | | | | | | | | | |
| - COMET | 80.71 | 92.44 | 79.02 | 88.90 | 88.15 | 97.32 | 87.76 | 79.80 | 91.95 | 79.54 | 87.11 | 89.30 | 94.54 | **87.04** |
| - COMETKIWI | 79.89 | 91.80 | 81.04 | 89.04 | 89.12 | 96.48 | **87.90** | 81.40 | 90.82 | 80.06 | 84.81 | 90.11 | 93.30 | 86.75 |
| **w/ External LLM-as-a-judge RM (Referenceless):** | | | | | | | | | | | | | | |
| - Qwen2.5-7B | 78.61 | 91.30 | 78.54 | 87.80 | 87.96 | 96.30 | 86.75 | 76.31 | 89.81 | 75.98 | 82.21 | 87.28 | 92.19 | 83.96 |
| - Qwen2.5-7B-Instruct | 79.10 | 91.58 | 79.28 | 88.56 | 87.98 | 96.19 | 87.12 | 77.03 | 89.73 | 76.60 | 82.16 | 87.87 | 92.07 | 84.24 |
| **w/ External LLM-as-a-judge RM (with Reference):** | | | | | | | | | | | | | | |
| - Qwen2.5-7B | 79.30 | 91.11 | 79.33 | 88.57 | 88.27 | 96.54 | 87.19 | 77.90 | 90.00 | 77.69 | 83.43 | 88.38 | 92.63 | 85.01 |
| - Qwen2.5-7B-Instruct | 79.10 | 91.58 | 79.28 | 88.56 | 87.98 | 96.19 | 87.12 | 77.03 | 89.73 | 76.60 | 82.16 | 87.87 | 92.07 | 84.24 |
| **Ours** | | | | | | | | | | | | | | |
| SSR-Zero-7B | 79.29 | 92.04 | 79.04 | 89.19 | 87.97 | 96.70 | 87.37 | 79.69 | 91.18 | 79.34 | 85.34 | 89.25 | 93.52 | 86.39 |
| - Ablation: w/ ref | 79.67 | 92.22 | 79.75 | 89.45 | 88.58 | 96.69 | 87.73 | 77.91 | 90.62 | 77.63 | 84.15 | 88.25 | 92.96 | 85.25 |
| SSR-X-Zero-7B | 80.62 | 91.92 | 80.56 | 89.42 | 88.84 | 96.62 | **88.00** | 81.11 | 91.56 | 79.67 | 86.75 | 90.08 | 93.98 | **87.19** |

Table 3: Translation quality of models trained via RL with different rewarding methods, measured by COMETKIWI-XXL (KIWI) and XCOMET-XXL (XCM) in English-Chinese directions (EN ↔ ZH). "Ablation: w/ ref" denotes a variant of SSR-Zero that includes reference translations in the judge prompt. **Bold and underlined** indicates the best-performing model, **bold only** the second-best, and underlined only the third-best.

Furthermore, SSR-Zero-7B substantially outperforms models with the same backbone trained using external LLM judges of the same size. This indicates that, during SSR training, improvements in translation capability may simultaneously enhance a model's judgment ability.

**Why does self-rewarding work?** An intuitive concern is that a model with weak translation capabilities should produce unreliable evaluation scores. However, we argue that the base model is not inherently weak in capability, but rather unaligned. As shown in Table 6, the backbone already demonstrates strong translation potential under few-shot prompting, confirming that the capability exists latently. SSR effectively unlocks this potential by using the model's internal judge to provide informative reward signals for RL training. Additionally, GRPO computes advantages via group-level reward normalization, making training depend on relative rankings within a batch rather than absolute scores, which provides robustness to noisy self-judgments. The convergence of SSR-Zero-3B (Table 1) – a smaller model with weaker judging capability – further supports this. Nonetheless, a quantitative diagnostic of judge–external metric disagreement remains a valuable direction for future work.

### 5.3 RQ3: Reference vs. Referenceless Rewarding

We further examine the influence of reference translations on reward signals and their subsequent impact on MT performance. Specifically, we introduce a variant of SSR-Zero that includes a reference translation in the judge prompt, denoted as "Ablation: w/ ref" in Table 3. The reference translation is obtained using the original target sentence from the training dataset. We use the same setting for LLM-as-a-judge baselines.

**Results** As shown in Table 3, the trained reference-based RM (COMET) and referenceless RM (COMETKIWI) yield similar results. For LLM-based external judges, explicitly providing reference translations typically leads to slightly higher performance compared to the reference-less setting. In self-reward training, the use of reference translations marginally improves performance in ZH → EN translation (from 87.37 to 87.73, +0.4%), but lowers the results for EN → ZH translation (from 86.39 to 85.25, -1.3%). In general, introducing reference translations to different reward methods does not consistently improve the model's performance, except when using external LLMs as judges. In particular, external references do not provide significant gains for SSR.

### 5.4 Additional Studies

**Multilingual Generalization** We evaluate SSR-Zero-7B (trained only on ∼13K EN↔ZH sen-

| Direction | Qwen2.5-7B-Inst. | SSR-Zero | Δ |
|---|---|---|---|
| ZH→X | 50.7 | **52.2** | +1.5 |
| X→ZH | 69.9 | **71.1** | +1.2 |
| EN→X | 50.7 | **53.6** | +2.9 |
| X→EN | **79.0** | 78.9 | -0.1 |
| X→Y (non-EN/ZH) | 38.9 | **42.2** | +3.3 |
| All | 41.7 | **44.9** | +3.2 |

Table 4: Multilingual evaluation on FLORES-200 (33 languages, XCOMET-XXL). SSR-Zero-7B is trained only on EN↔ZH.

| Direction | Model | Mean DA | z-score | Std |
|---|---|---|---|---|
| EN→ZH | SSR-Zero-7B | **85.65** | +0.073 | ±0.90 |
| | Qwen2.5-7B-Inst. | 82.90 | -0.073 | ±1.00 |
| ZH→EN | SSR-Zero-7B | **91.05** | +0.050 | ±0.80 |
| | Qwen2.5-7B-Inst. | 89.69 | -0.050 | ±0.97 |

Table 5: Human Direct Assessment on WMT23 EN↔ZH.

tences) on FLORES-200 covering 33 languages using XCOMET-XXL (Table 4). Despite never seeing other language pairs during training, SSR-Zero-7B improves translation quality across diverse directions, including X→Y pairs not involving English or Chinese (+3.3 on average). We attribute this to shared cross-lingual representations in the pretrained backbone and the universal nature of the quality criteria learned through self-judging. Representative per-pair results are in Appendix A.2. That said, training remains centered on EN↔ZH, and generalization claims should be interpreted accordingly.

**Human Evaluation** We conduct a segment-level Direct Assessment (DA) on WMT23 EN↔ZH (200 samples per direction, two bilingual annotators). Scores were z-normalized per annotator following WMT protocols. As shown in Table 5, SSR-Zero-7B achieves higher mean DA scores than Qwen2.5-7B-Instruct in both directions, confirming that automatic metric gains correspond to real improvements in perceived quality.

**Comparison with Few-Shot Prompting** We compare SSR-Zero-7B with few-shot prompting baselines (0/1/5/10-shot) using the same backbone (Table 6). Few-shot prompting improves the backbone but saturates after one shot, reaching a level comparable to Qwen2.5-7B-Instruct. SSR-Zero-7B substantially outperforms all variants without any in-context examples, demonstrating that the gains from self-rewarding RL training cannot be replicated by simply providing parallel examples at inference time.

| Model | ZH→EN | EN→ZH |
|---|---|---|
| Qwen2.5-7B (0-shot) | 79.52 | 79.06 |
| Qwen2.5-7B (1-shot) | 89.39 | 84.37 |
| Qwen2.5-7B (5-shot) | 89.29 | 84.82 |
| Qwen2.5-7B (10-shot) | 89.36 | 84.75 |
| Qwen2.5-7B-Instruct | 90.19 | 86.31 |
| SSR-Zero-7B (0-shot) | **92.64** | **90.01** |

Table 6: Average XCOMET-XXL scores for few-shot baselines vs. SSR-Zero-7B.

## 6 Conclusion

In this work, we propose 💫**SSR**, a simple yet effective reinforcement learning approach for machine translation. SSR does not rely on external reward models (RMs) or reference data; instead, it leverages the actor model itself as a judge to generate reward signals and optimize its performance through GRPO training. Initialized from an uninstructed Qwen2.5-7B backbone, our SSR-Zero-7B model outperforms many open-source MT-specific LLMs, such as TowerInstruct-13B, as well as larger general-purpose LLMs like Qwen2.5-32B-Instruct across multiple English ↔ Chinese translation benchmarks.

Our analysis shows that SSR is more effective than using same-size external LLM-as-a-judge models. In addition to high-resource settings, SSR demonstrates strong generalization to low-resource language pairs, yielding substantial improvements when trained with limited monolingual data. Although SSR alone slightly underperforms dedicated RMs (e.g., COMET and COMETKIWI) trained on large-scale annotated MT-evaluation data, combining SSR with these RMs yields consistent additional improvements. Our best-performing model, SSR-X-Zero-7B, integrates SSR with COMET and achieves competitive performance relative to existing open-source and closed-source systems on English ↔ Chinese translation benchmarks.

These findings provide insight into reward selection for MT via reinforcement learning and highlight that strong pre-trained LLMs inherently possess reliable MT evaluation capabilities that can be exploited to improve translation quality. Overall, our work demonstrates the potential of self-reward-based RL approaches to reduce dependence on costly external supervision from humans or trained reward models, particularly in low-resource scenarios.

## Limitations

Our work demonstrates the effectiveness and, to some extent, the generalizability of self-reward training for machine translation. However, the applicability of this approach across different model architectures remains unexplored. Prior work has shown that R1-Zero-like training can exhibit varying effectiveness across model families (Gandhi et al., 2025), and it therefore remains unclear whether SSR can consistently incentivize strong MT capabilities in architectures such as Llama.

Although our few-shot comparison (Table 6) shows that providing in-context examples to the actor does not match SSR's gains, the effect of alternative prompting strategies for the judge – such as Chain-of-Thought (CoT) or few-shot prompting – remains unexplored. That said, recent work by Qian et al. (2024) suggests that CoT and 5-shot prompting do not outperform zero-shot prompting for MT evaluation when using 7B-scale models with similar evaluation prompts.

In addition, using the model itself as a judge may introduce a potential risk of reinforcing incorrect self-judgments or exhibiting reward bias. As discussed in Section 5.2, our evidence suggests that GRPO's group-level normalization and the backbone's latent capabilities provide sufficient robustness, and the qualitative analysis in Appendix A.3 shows that the judge can identify major errors despite occasional misjudgments. Nevertheless, a deeper theoretical and empirical characterization of failure modes in self-rewarding remains necessary, particularly for weaker backbone models.

Another limitation is that our evaluation relies heavily on automatic metrics rather than direct human assessment. We use XCOMET-XXL and COMETKIWI-XXL, which are recommended by the WMT community[6] and have been shown to correlate strongly with human judgments. Although our small-scale human evaluation (Table 5) confirms that metric gains correspond to real quality improvements, a larger-scale human assessment covering more systems and language pairs remains an important direction for future work.

Furthermore, our training is centered on EN↔ZH. While Section 5.4 demonstrates encouraging cross-lingual transfer to 33 unseen languages, systematic evaluation across more diverse training language pairs is needed to fully validate the generalizability of SSR.

Finally, recent studies (Liu et al., 2025) indicate that LLM-as-a-judge frameworks may benefit from test-time scaling techniques such as voting. We leave the exploration of such techniques within the context of SSR-based training to future work.

[6] https://www2.statmt.org/wmt25/translation-task.html

## References


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

Anthropic. 2024. [link].

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, and 1 others. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive

behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.

Xiang Geng, Ming Zhu, Jiahuan Li, Zhejian Lai, Wei Zou, Shuaijie She, Jiaxin Guo, Xiaofeng Zhao, Yinglu Li, Yuang Li, and 1 others. 2024. Why not transform chat large language models to non-english? *arXiv preprint arXiv:2405.13923*.

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

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

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Yan Meng, Di Wu, and Christof Monz. 2024. How to learn in a noisy world? self-correcting the real-world data noise on machine translation. *arXiv preprint arXiv:2407.02208*.

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

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orăsan, Tharindu Ranasinghe, and Frédéric Blain. 2024. What do large language models need for machine translation evaluation? *arXiv preprint arXiv:2410.03278*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

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.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© 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.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024a. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, José Pombal, Nuno M Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, and 1 others. 2024b. Tower v2: Unbabel-ist 2024 submission for the general mt shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan

Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Mingyang Song, Mao Zheng, Zheng Li, Wenjie Yang, Xuan Luo, Yue Pan, and Feng Zhang. 2025. Fastcurl: Curriculum reinforcement learning with progressive context extension for efficient training r1-like reasoning models. *arXiv preprint arXiv:2503.17287*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775.

K Uhlig, J Wuebker, R Reinauer, and J DeNero. 2025. Cross-lingual human-preference alignment for neural machine translation with direct quality optimization. In *Proceedings of the Tenth Conference on Machine Translation*.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2024. Drt: Deep reasoning translation via long chain-of-thought. *arXiv preprint arXiv:2412.17498*.

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

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Lijun Wu, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2017. Sequence prediction with unlabeled data by reward function learning. In *IJCAI*, pages 3098–3104.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Nuo Xu, Jun Zhao, Can Zu, Sixian Li, Lu Chen, Zhihao Zhang, Rui Zheng, Shihan Dou, Wenjuan Qin, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Advancing translation preference modeling with rlhf: A step towards cost-effective solution. *arXiv preprint arXiv:2410.07515*.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2.5 technical report. *arXiv e-prints*, pages arXiv–2412.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2024b. Language imbalance driven rewarding for multilingual self-improving. *arXiv preprint arXiv:2410.08964*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025a. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*.

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. 2025b. Process-based self-rewarding language models. *ArXiv*, abs/2503.03746.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

Wei Zou, Sen Yang, Yu Bao, Shujian Huang, Jiajun Chen, and Shanbo Cheng. 2025. Trans-zero: Self-play incentivizes large language models for multilingual translation without parallel data. *arXiv preprint arXiv:2504.14669*.

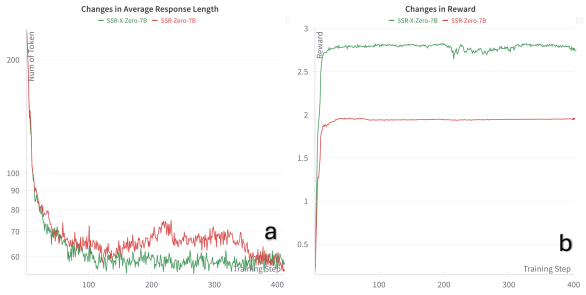


Figure 2: Changes in average response length (a) and training rewards (b) of SSR/SSR-X-Zero-7B during GRPO training.

# A Appendix

## A.1 Training Dynamics of SSR

We also report how the response length and test set performance evolve during SSR/SSR-X-Zero-7B training. As shown in Fig. 2, we did not observe the increase in output length typical of R1-like training in mathematics (Guo et al., 2025), nor the curve seen in Feng et al. (2025) which first decreases and then increases. As training progressed, the model quickly reduced the output length from about 200 to 60-70 tokens and did not generate meaningful CoTs. A typical CoT before translation was "<think> *I need to translate this sentence from {src_lang} to {tgt_lang}.*</think>".

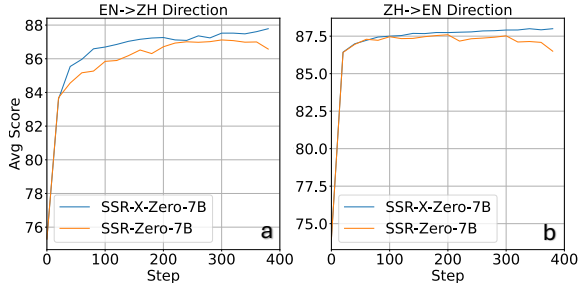


Figure 3: Changes in translation quality during training, measured by the average scores of COMETKIWI-XXL and XCOMET-XXL on the EN → ZH (a) and ZH → EN (b) benchmarks.

Despite this, we observed an increase trend in performance in the test set as training progressed, as shown in Figure 3. We also noticed that the performance of SSR-Zero-7B for EN → ZH saturates after approximately 3 epochs (around 300 steps) and decreases afterward, while its ZH → EN performance converges earlier, at roughly 200 steps. In contrast, SSR-Zero-X-7B demonstrates better stability and continuous improvement during training. Upon inspection, we found that SSR-Zero-7B began enclosing translated outputs with extraneous quotation marks (i.e., <answer>"translated text"</answer>) after 300 steps, which our regular expression could not filter out during evaluation. This formatting issue led automated metrics XCOMET-XXL and COMETKIWI-XXL to produce lower evaluation scores. This issue was not observed during the SSR-Zero-X-7B's training. We leave further exploration in maintaining consistent output formatting of SSR training for future work.

## A.2 FLORES-200 Per-Pair Results

Table 7 shows representative per-pair results from our FLORES-200 evaluation (33 languages) using XCOMET-XXL. SSR-Zero-7B is trained only on EN↔ZH data. **Bold** indicates the better score between Qwen2.5-7B-Instruct and SSR-Zero-7B.

| Direction | Qwen2.5-7B-Inst. | SSR-Zero |
|---|---|---|
| EN→DE | 90.4 | **93.1** |
| EN→ES | 88.3 | **93.4** |
| EN→FR | 82.7 | **89.6** |
| EN→JA | 67.5 | **78.5** |
| EN→KO | 50.8 | **61.3** |
| EN→VI | 73.2 | **81.2** |
| EN→RU | 72.7 | **79.4** |
| EN→PT | 88.1 | **92.0** |
| DE→EN | 94.5 | **95.4** |
| JA→EN | 90.0 | **91.4** |
| KO→EN | 88.7 | **90.9** |
| ID→EN | 92.8 | **93.4** |
| ZH→DE | 80.9 | **88.1** |
| ZH→ES | 82.6 | **90.4** |
| ZH→KO | 56.8 | **68.4** |
| ZH→RU | 73.5 | **83.1** |
| ZH→MS | 68.0 | **76.9** |
| RU→ZH | 87.0 | **88.9** |
| ES→ZH | 88.2 | **91.2** |
| IT→ZH | 83.1 | **86.6** |
| PT→ZH | 84.7 | **89.1** |

Table 7: Representative per-pair XCOMET-XXL scores on FLORES-200. SSR-Zero-7B is trained only on EN↔ZH data.

## A.3 Qualitative Analysis of Self-Judging

We present qualitative examples illustrating (a) translation improvements from backbone to SSR-Zero, (b) cases where the self-judge produces accurate assessments, and (c) cases where it misjudges.

**Translation Improvements Case 1 (Subject misinterpretation):**

- **Source:** "我听说她被艾尔文王后附身了，所以总是去那个房间。"
- **Qwen2.5-7B:** "I heard she was possessed by Queen Elvina, so **I** always go to that room." (wrong subject)

- **SSR-Zero-7B:** "I heard that she was possessed by Queen Elvin, which is why **she** always went to that room." ✓

**Case 2 (Literal translation error):**

- **Source:** "创新高了。"
- **Qwen2.5-7B:** "**Innovations peaked.**" (confuses "创新高/hit a new high" with "创新/innovation")
- **SSR-Zero-7B:** "**It set a new record.**" ✓

**Accurate Self-Judgments Case 1 (Detecting name mistranslation):**

- **Source:** "In the fourth set, Zhu Yuling made a good start and was 4-2 ahead."
- **Actor:** "第四局, 荣钰婷开局不错, 导致了4-2的领先。"
- **Judge:** "The Chinese translation mentions '荣钰婷' instead of 'Zhu Yuling' and adds '导致了 (caused)' not in the original." → Score: 70

**Case 2 (Detecting repetition):**

- **Source:** "军训日晒时间较长，防晒霜每隔两小时擦一次才能有持久效果，千万不要偷懒。"
- **Actor:** "...never slack off **slacking off**."
- **Judge:** "It contains repetitive content such as 'slacking off' and 'never slack off,' and uses informal language." → Score: 70

**Inaccurate Self-Judgments Case 1 (Overlooking fluency):**

- **Source:** "ABC was second with 3.26 million viewers."
- **Actor:** "ABC获得了326万观众的第二名。" (awkward; more natural: "ABC以326万观众位居第二")
- **Judge:** "The grammar and meaning are preserved." → Score: 100 (expected ∼85)

**Case 2 (Hallucinating an error):**

- **Source:** "美国乘客Bill Jaynes称，这架飞机飞得很低。"
- **Actor:** "**American** passenger Bill Jaynes said that the plane was flying very low."
- **Judge:** "It is missing the name of the country where the passengers are from." → Score: 80 (expected 100)

These examples show that the self-judge can identify major errors such as name mistranslation and repetition, yet may overlook fluency issues or hallucinate errors. Despite these imperfections, the self-judge provides a sufficiently robust reward signal for RL training, as evidenced by consistent improvements on both automatic metrics and human evaluation.