FeedTrans: Enhance Machine Translation via Feedback

Zengkui Sun[1]*, **Jiali Zeng**[2], **Jiaan Wang**[2], **Fandong Meng**[2], **Xinyan Guan**[2], **Yufeng Chen**[1], **Jinan Xu**[1], **Wenjuan Han**[1]†, and **Jie Zhou**[2]
[1]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University, Beijing, China
[2]WeChat AI, Tencent Inc, China
{zengksun,chenyf,jaxu,wjhan}@bjtu.edu.cn, guanxy0406@gmail.com
{lemonzeng,torchwang,fandongmeng,withtomzhou}@tencent.com

Abstract

Large reasoning models (LRMs) have shown exceptional performance in complex tasks such as mathematics and coding. In the field of machine translation (MT), reinforcement learning (RL) has been utilized to enhance the quality of translations. However, traditional RL approaches rely heavily on the base model's inherent translation capabilities, which may falter when dealing with terminology translations and domain-specific expressions without sufficient guidance. In this paper, we introduce **FeedTrans** (**Feed**back-driven **Trans**lation), which employs a *FeedRollout* mechanism to incorporate feedback as guidance, enabling the production of high-quality translations and expanding the search space for improved translation outcomes. Extensive experiments across six translation tasks validate the effectiveness of our approach. To the best of our knowledge, this is the first attempt at utilizing feedback in MT-oriented RL. https://github.com/Acerkoo/FeedTrans

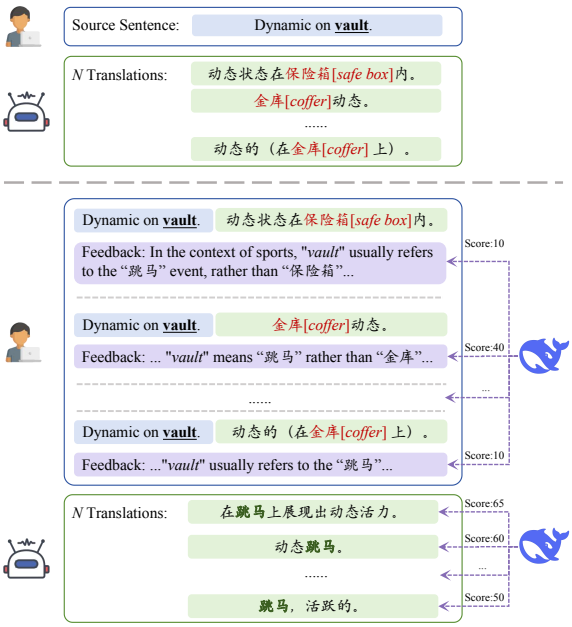


Figure 1: Example of terms in translation. The ground truth is "跳马时充满活力", where "vault" denotes a gymnastics event "vaulting horse", *i.e.*, "跳马", rather than "coffer" or "safe box".

1 Introduction

Large reasoning models (LRMs) (Jaech et al., 2024; Guo et al., 2025a) have demonstrated remarkable performance in complex tasks like mathematics (Chen et al., 2025b; Li et al., 2025b), coding (Zhang et al., 2024; Su et al., 2025), and question generation (Guan et al., 2025). In machine translation (MT), Marco-o1 (Zhao et al., 2024) and DRT-o1 (Wang et al., 2024a) fine-tune LRMs on synthesized long-thought data for translation. More recently, due to the robust generalization capabilities of RL (Guo et al., 2025a; Luo et al., 2025; Yu et al., 2025), researchers are increasingly applying RL to translation tasks (He et al., 2025; Feng et al., 2025b), using COMET (Rei et al., 2022a) score or large language model (LLM) like DeepSeek-v3 (Liu et al., 2024) as an evaluator to provide reward signals for the RL training.

Despite their promising results, two challenges remain: 1) Existing RL methods are highly reliant on the base model's translation capabilities (Yan et al., 2025; Zhao et al., 2025; Yue et al., 2025). However, these models often struggle to produce satisfactory translations for specialized terminology and domain-specific expressions without additional guidance, thereby limiting the potential of RL to enhance translation quality. 2) When using an LLM as the evaluator, it not only scores translations but also provides reasons and feedback. However, previous studies have primarily focused on designing evaluation rules (Wang et al., 2025b),

*Work was done when Sun was interning at WeChat AI, Tencent Inc, China.
†Wenjuan Han is the corresponding author.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1899–1916, 2026. https://doi.org/10.1162/TACL.a.784
Action Editor: Barry Haddow. Submission batch: 6/2025; Revision batch: 2/2026; Published 8/2026.

leaving the use of feedback to enhance RL efficacy largely unexplored.

In this paper, we introduce **FeedTrans** (**Feed**back-driven **Trans**lation), which leverages feedback to enhance the exploration capabilities of LRMs during RL. We propose a novel *FeedRollout* mechanism. Specifically, given a source input, we first sample N translations and use an LLM (*e.g.*, Deepseek-v3) as the evaluator to provide textual feedback and a translation quality score. For translations deemed low-quality, we integrate the feedback into standard reasoning trajectories and perform an additional rollout process to refine the translation. As shown in Figure 1, for the term "vault", the translations sampled by the base model are all incorrect. Despite slight variations in their scores, none accurately translate to "跳马" (vaulting horse). With guidance from DeepSeek-v3 feedback, the model successfully samples the correct translation and incorporates it into the RL training process. In summary, FeedTrans enables LRMs to engage in more thoughtful reasoning with external guidance, leading to higher-quality translations and expanding their search space for more efficient and effective MT-oriented RL.

We conducted extensive experiments across six translation tasks, including two in-domain and four out-of-domain tasks. The experimental results demonstrate that our FeedTrans-7B significantly outperforms baseline models ranging from 7B to 14B. Additionally, we performed an in-depth analysis to uncover the impact of the *FeedRollout* mechanism. Interestingly, our model, trained using FeedRollout, developed enhanced reflective capabilities, further improving translation performance.

2 Related Work

2.1 Deep Reasoning Machine Translation

Recently, Large Reasoning Models (Jaech et al., 2024; Guo et al., 2025a; Team et al., 2025) have achieved remarkable performance on complex reasoning tasks by generating long thought (Su et al., 2025; Zhang et al., 2024; Luo et al., 2025; Hu et al., 2025; Wang et al., 2024a; Guan et al., 2025). Chen et al. (2025b) provide a systematic survey on deep reasoning LLMs to introduce the development process. More recently, researchers scale the capability of deep reasoning to machine translation (Zhao et al., 2024; Wang et al., 2024a). Marco-o1 (Zhao et al., 2024), DRT-o1 (Wang et al., 2024a), and SlangDIT (Liang et al., 2025) explore the application of LRMs in MT, which supervised fine-tuning (SFT) models on the synthesized long-thought data. Further, empirical studies (Chen et al., 2025a; Liu et al., 2025a) showcase the superior performance of LRMs on translation tasks, especially in historical and cultural translation. Moreover, building upon the success of RL (Guo et al., 2025a; Luo et al., 2025; Yu et al., 2025), researchers are increasingly applying RL to translation tasks (He et al., 2025; Feng et al., 2025b). Specifically, R1-T1 (He et al., 2025) and MT-R1-Zero (Feng et al., 2025b) employ the COMET (Rei et al., 2022a) or COMETKiwi (Rei et al., 2022b) score as the reward in RL. Additionally, DeepTrans (Wang et al., 2025b,c) and SSR-Zero (Yang et al., 2025) design the reward from the external LLMs like DeepSeek-v3 (Liu et al., 2024) or the policy model to be self-rewarding. However, the above approaches are highly reliant on the base model's translation capabilities (Yan et al., 2025; Zhao et al., 2025; Yue et al., 2025). Without additional guidance, these models often struggle to produce satisfactory translations for specialized terminology and domain-specific expressions. Compared to them, we introduce real-time feedback to the errors in the generated translation during RL.

2.2 Feedback in Large Reasoning Models

With the emergence of RL, Guo et al. (2025a) demonstrate that direct RL could foster the self-reflection mechanism without supervision, *e.g.*, "Aha moment". Meanwhile, Liu et al. (2025b) observe that self-reflection does not necessarily lead to the correct answer, and the performance of RL is bounded by the base LLM itself (Yan et al., 2025; Zhao et al., 2025; Yue et al., 2025). In addition, researchers (Li et al., 2025a; Feng et al., 2025a; Qian et al., 2025; Wang et al., 2025a) reveal that the thought process with the feedback from the integrated external tools could assist the LRMs in escaping from mistakes with reflection and solving problems more accurately. Before the emergence of LRMs, several studies (Chen et al., 2023; Raunak et al., 2023; Chen et al., 2024; Wang et al., 2024b) propose to facilitate a similar feedback and reflection process, utilizing multi-step inference. However, the utilization of feedback in the thought process to enhance the translation quality of LRMs is underexplored.

A translation question requires translating a given text from {*src_lang*} into {*tgt_lang*}. PROMPT to DeepSeek-v3

The given text is as follows:
<text>
{*source*}
</text>

Someone did this translation question, provide the translation of the given text:
<translation>
{*trans*}
</translation>

Please critically evaluate the {*tgt_lang*} translation of the given {*src_lang*} text.

Rate the translation on a scale of 0 to 100, where:
- 10 points: Poor translation; the text is somewhat understandable but contains significant errors.
- 30 points: Fair translation; the text conveys the basic meaning but lacks fluency and contains several awkward phrases or inaccuracies.
- 50 points: Good translation; the text is mostly fluent and conveys the original meaning well, but may have minor awkwardness or slight inaccuracies.
- 70 points: Very good translation; the text is smooth and natural, effectively conveying the intended meaning, but may still have minor issues that could slightly affect understanding.
- 90 points: Excellent translation; the text is fluent and natural, conveying the original meaning clearly and effectively, with no significant issues that would hinder understanding.

Please evaluate the translation and enumerate the existing key errors in the translation in brief. Besides, determine the score for the translation.

For the evaluated result, format it as below:
Key Errors in Translation
{*Feedback*}
Translation Score
{"*Score*": [int]}

Figure 2: Prompt DeepSeek-v3 to evaluate the translation quality and provide feedback during RL. The "*src_lang*", "*tgt_lang*", "*source*", and "*trans*" denote the source language, target language, the source text, and the generated translation by the policy model, respectively. "*Feedback*" and "*Score*" denote the evaluation result of DeepSeek-v3, respectively.

3 FeedTrans

In this section, we first introduce the reward design of FeedTrans (§ 3.1). In addition to quantifying the translation quality, the reward model is requested to provide translation feedback with textual suggestions and critiques. Then, we design a two-stage rollout process, named FeedRollout, to enhance the model translations based on the translation feedback (§ 3.2).

3.1 Reward Design

Similar to the previous studies (Guo et al., 2025a), FeedTrans first thinks and analyzes how to translate the sentence, followed by the translation result.

The format of model generation is defined as "`<think>[thought]</think>[translation]`", where "`[thought]`" and "`[translation]`" denote the content of thought and final translation, respectively. "`<think>`" and "`</think>`" are two special tokens to indicate the boundary of thought content. Building upon this format, we design the following rewards:

Format Reward. To let FeedTrans generate the correct format, we employ a regular expression to judge whether the model generation meets our definition:

$$r_{\text{format}} = \begin{cases} 1 & \text{if format is correct,} \\ 0 & \text{others} \end{cases} \quad (1)$$

The format reward ensures that FeedTrans generates the correct format as its foundation.

Translation Reward. In view of the strong ability of LLM-as-a-judge (Wang et al., 2023; Huang et al., 2025), we employ DeepSeek-v3 (Liu et al., 2024) (671B) as the reward model. Different from previous studies (Wang et al., 2025b; He et al., 2025; Feng et al., 2025b), which only quantify the translation quality, we make the reward model provide quantified rewards and textual translation feedback simultaneously. As shown in Figure 2, we design an evaluation prompt to enable the reward model to identify translation errors and quantify the translation quality:

$$f_{\text{trans}}, r_{\text{trans}} = \text{DS-V3}^{\text{rm}}(source, trans), \quad (2)$$

where DS-V3^{rm} denotes using DeepSeek-v3 as the reward model. r_{trans} denotes the translation reward on a 100-point scale; while f_{trans} indicates the translation feedback with textual suggestions and critiques.

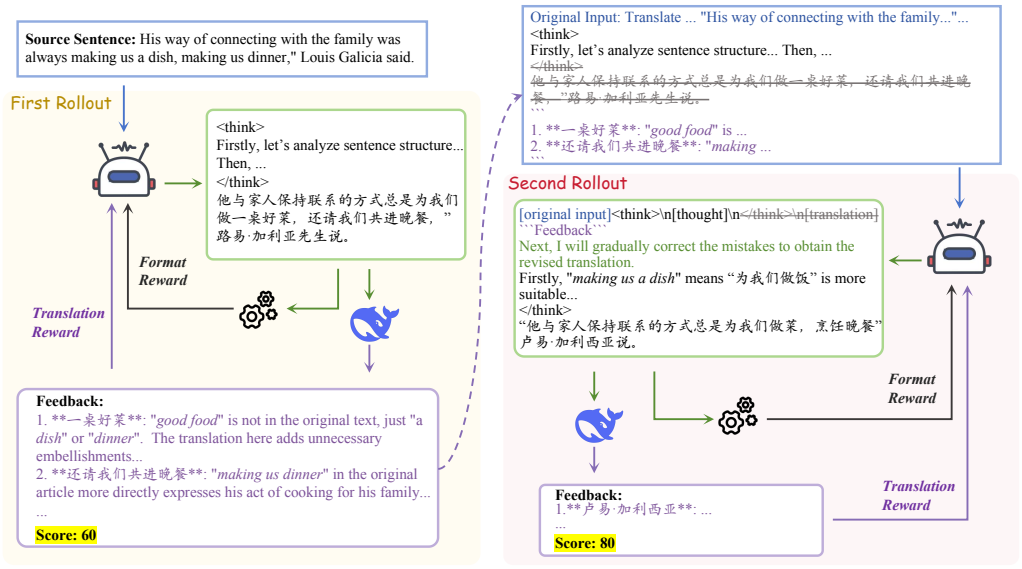


Figure 3: The overview of the FeedRollout process, containing the first and second rollout stages. The policy model first samples responses and obtains the format and translation reward. For the flawed translation (*e.g.*, Score ≤ 60), the policy model injects the feedback into thought to explore better translation.

Overall Reward. Combining the above two types of reward, the final reward is as follows:

$$r = \begin{cases} r_{\text{trans}} & \text{if } r_{\text{format}} = 1, \\ 0 & \text{others} \end{cases} \tag{3}$$

3.2 FeedRollout

As we described above, in addition to the translation reward (r_{trans}), the reward model also provides valuable translation feedback f_{trans}. Therefore, we explore whether the feedback can further enhance the translations. Given this motivation, we design a novel rollout process, *i.e.*, FeedRollout, which involves two rollout stages to guide the translation model with feedback, illustrated in Figure 3.

First Rollout. In the first rollout stage, the translation model conducts a vanilla rollout process for each source sentence, *i.e.*, sampling N responses with deep reasoning for translation, which we denote as $T = \{t_1, t_2, \ldots, t_N\}$. Consequently, the reward model (DeepSeek-v3) provides the corresponding evaluation, with the translation feedback $F = \{f_1, f_2, \ldots, f_N\}$ and translation scores $S = \{s_1, s_2, \ldots, s_N\}$. f_i and s_i denote the translation feedback and translation reward of t_i, respectively. Based on S, we extract flawed translations:

$$\text{Flawed}(t_i) = \begin{cases} \text{False} & s_i > \tau, \\ \text{True} & s_i \leq \tau \end{cases} \tag{4}$$

where $\text{Flawed}(t_i)$ denotes whether t_i is a flawed translation. τ is a pre-defined threshold. In practice, we treat the translations with scores in the bottom τ-% in batch as flawed translations, to maintain a stable number of flawed translations.

Second Rollout. For flawed translations, we next guide the translation model on how to refine them in the second rollout process. We speculate that flawed translation may be caused by the base models' translation capabilities (Yan et al., 2025; Zhao et al., 2025; Yue et al., 2025). The translation feedback f_i might assist the thoughts and guide the model to provide better translations beyond the initial translation capabilities. Thus, for each flawed translation, we concatenate the translation feedback after the original thought content, and let the model re-generate the translation, like the blue box on the upper right in Figure 3. Then, we insert one sentence "*Next, I will gradually correct the mistakes to obtain the revised translation.*" to simulate the continual step in thought and guide the model to refine the flawed translation. Building upon the second rollout, we also calculate its reward as we discussed in Eq. 3. In this way, the policy model is encouraged to explore better translations under the guidance of translation feedback.

3.3 Training Process

To train FeedTrans, we first leverage cold-start SFT to encourage the model with a "think-then-translate" pattern. Following Wang et al. (2025b), we select 4K general-domain English sentences

from WMT24, and employ DeepSeek-R1 (Guo et al., 2025a) to perform translations. The generated thoughts and translations are used to SFT our ColdStart model.

Next, we employ RL training to further enhance the model. Similar to previous approaches (He et al., 2025; Wang et al., 2025b; Feng et al., 2025b), we use the GRPO algorithm (Shao et al., 2024) to optimize the model. In detail, for a source query q, the behavior policy π_{old} firstly samples a group of N individual responses $\{t_i\}_{i=1}^{N}$. Consequently, π_{old} samples an extra group of N responses $\{t_j\}_{j=1}^{N}$, including the reused rewards of good translations and new rewards for the generated translations in the second rollout. Then, we merge both groups to calculate the advantage, where the advantage of the i-th response is calculated by normalizing the group-level rewards $\{r_i\}_{i=1}^{2N}$:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{2N})}{\text{std}(\{r_i\}_{i=1}^{2N})}. \tag{5}$$

Finally, the GRPO optimizes the policy model π_{new} by maximizing the following objective:

$$\begin{aligned}\mathcal{J}_\pi = \frac{1}{2N}\sum_{1}^{2N}(& \min(\nabla_\pi \hat{A}_i, \\ & \text{clip}(\nabla_\pi, 1-\epsilon, 1+\epsilon)\hat{A}_i) \\ & - \beta \mathcal{KL}(\pi_{new}||\pi_{ref}),\end{aligned} \tag{6}$$

$$\nabla_\pi = \frac{\pi_{new}(t_i|q)}{\pi_{old}(t_i|q)}, \tag{7}$$

where ϵ and β are hyperparameters, π_{ref} is the reference model. Following DAPO (Yu et al., 2025), we remove the $\mathcal{KL}$ item in the GRPO algorithm, *i.e.*, $\beta = 0$. To prevent the model from attempting to memorize specific "*{Feedback}*", we mask the corresponding tokens when calculating the RL objectives, similar to Li et al. (2025a).

4 Experiments

4.1 Experimental Setups

Dataset. As described in § 3.3, our training process contains the ColdStart and RL stages. For the ColdStart stage, we prepare the training sets in four translation directions, *i.e.*, English (En) ↔ Chinese (Zh) and English (En) ↔ German (De). To create this training set, we select 4K general-domain English sentences from WMT24 and employ DeepSeek-R1 to translate them into Chinese and German, under the 'think-then-translate' pattern. We then back-translate the outputs to create Chinese / German → English training data. For the RL stage, we collect parallel sentences from the WMT 2017-2020 dev sets of four translation directions, with roughly 10K sentences (in news domain) per direction. Additionally, we also employ DeepSeek-R1 as the teacher model to translate the source sentence within the training sets of RL stage. To evaluate our models, we use Flores-200 (Goyal et al., 2022; Costa-Jussà et al., 2022), which covers domains such as news, children's books, and travel guides, to assess translation performance in four directions. Further, in case of English-to-Chinese translation, we investigate the translation ability of multi-domain to evaluate the generalization of MT models after tuning. In this multi-domain scenario, we utilize test sets from CultureMT (Yao et al., 2023), Tico19 (Anastasopoulos et al., 2020), MetaphorTrans (Wang et al., 2024a), and RedTrans (Guo et al., 2025b), focusing on the cultural, medical, literary, and social network service domains.

Metrics. We assess translation quality using two widely used complementary metrics: the reference-based semantic metric COMET (Rei et al., 2020, 2022a) (wmt22-comet-da), and the reference-free semantic metric COMETKiwi (Rei et al., 2022b) (wmt22-cometkiwi-da).

Backbones. We select Qwen2.5-7B-Instruct (Yang et al., 2024) as our backbone for fair comparison, following previous studies (Wang et al., 2024a, 2025b; Feng et al., 2025b). Besides, we add Llama3-8B-Instruct (Grattafiori et al., 2024) to evaluate our method with a different model architecture. Moreover, we add the Qwen2.5-14B-Instruct as the larger backbone.

4.2 Baselines

(1) *General Non-reasoning LLMs*. We use Llama3-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct (Yang et al., 2024) as baselines. Building upon the above three models, we conduct SFT with the golden parallel sentences and the distilled version from DeepSeek-v3 for the comprehensive comparison, and use "*-Golden*" and "*-V3*" as their corresponding suffixes.

(2) *General LRMs*. Marco-o1-7B (Zhao et al., 2024), DeepSeek-R1-Distill-Qwen-7B (R1-Qwen-7B), DeepSeek-R1-Distill-Llama-8B (R1-Llama-8B), and DeepSeek-R1-Distill-Qwen-14B (R1-Qwen-14B) (Guo et al., 2025a) are used as base-

Model	CoT?	En→Zh	Zh→En	En→De	De→En
DeepSeek-v3	×	89.87 / 86.22	87.74 / 85.98	89.16 / 86.98	89.67 / 86.03
Backbone: Qwen2.5-7B-Instruct					
Qwen2.5-7B-Instruct	×	87.75 / 84.36	86.35 / 84.38	82.75 / 80.85	88.06 / 85.19
R1-Qwen-7B	✓	83.83 / 79.53	83.91 / 82.53	78.31 / 75.90	84.39 / 81.19
Macro-o1	✓	87.51 / 84.40	86.44 / 85.18	83.98 / 82.54	88.46 / 85.35
DRT-7B	✓	86.99 / 84.38	–	–	–
DeepTrans-7B	✓	87.40 / 85.33	–	–	–
MT-7B-*ColdStart* (w/o CoT)	×	86.67 / 82.53	86.12 / 84.77	84.21 / 82.30	88.24 / 85.26
MT-7B-*FullData* (w/o CoT)	×	87.97 / 84.87	86.62 / 85.30	85.19 / 83.53	88.56 / 85.41
MT-7B-*ColdStart* (w/ CoT)	✓	86.82 / 82.97	86.17 / 84.81	84.82 / 83.11	88.31 / 85.56
MT-7B-*FullData* (w/ CoT)	✓	88.12 / 84.96	86.73 / 85.41	85.34 / 83.90	88.66 / 85.65
FeedTrans-7B (***Ours***)	✓	**88.81 / 85.89**	**86.89 / 85.47**	**85.88 / 84.47**	**88.84 / 85.70**
w/o FeedRollout	✓	88.34 / 85.42	86.71 / 85.39	85.28 / 84.06	88.74 / 85.63
Backbone: Llama3-8B-Instruct					
Llama3-8B-Instruct	×	84.97 / 81.63	86.18 / 84.41	85.90 / 84.17	88.22 / 85.39
R1-Llama-8B	✓	85.19 / 81.87	84.87 / 83.65	76.85 / 75.57	87.20 / 84.30
MT-8B-*ColdStart* (w/o CoT)	×	86.68 / 83.89	85.32 / 84.31	85.84 / 84.12	88.32 / 85.41
MT-8B-*FullData* (w/o CoT)	×	87.26 / 84.59	86.31 / 85.07	86.28 / 84.18	88.57 / 85.58
MT-8B-*ColdStart* (w/ CoT)	✓	86.89 / 84.09	85.58 / 84.36	86.35 / 84.69	88.27 / 85.49
MT-8B-*FullData* (w/ CoT)	✓	87.45 / 84.69	86.43 / 85.12	86.67 / 84.87	88.69 / 85.59
FeedTrans-8B (***Ours***)	✓	**87.97 / 85.11**	**86.62 / 85.38**	**87.06 / 85.05**	**88.93 / 85.75**
w/o FeedRollout	✓	87.53 / 84.75	86.35 / 85.20	86.78 / 84.74	88.89 / 85.69
Backbone: Qwen2.5-14B-Instruct					
Qwen2.5-14B-Instruct	×	88.08 / 84.59	87.11 / 85.05	83.34 / 81.90	89.11 / 85.45
R1-Qwen-14B	✓	87.94 / 84.74	86.53 / 85.33	84.19 / 83.36	88.81 / 85.68
DRT-14B	✓	87.19 / 84.59	–	–	–
MT-14B-*ColdStart* (w/o CoT)	×	88.11 / 85.08	86.89 / 85.12	86.37 / 84.65	88.11 / 85.08
MT-14B-*FullData* (w/o CoT)	×	88.56 / 85.62	87.13 / 85.59	86.94 / 85.00	88.70 / 85.51
MT-14B-*ColdStart* (w/ CoT)	✓	88.15 / 85.07	86.44 / 85.26	86.87 / 85.23	88.32 / 85.33
MT-14B-*FullData* (w/ CoT)	✓	88.64 / 85.57	87.22 / 85.64	87.34 / 85.68	89.01 / 85.62
FeedTrans-14B (***Ours***)	✓	**89.11 / 86.15**	**87.31 / 85.74**	**87.59 / 85.97**	**89.27 / 85.94**
w/o FeedRollout	✓	88.75 / 85.82	87.09 / 85.48	87.17 / 85.60	88.99 / 85.78

Table 1: Translation results in four translation directions of the Flores test set. 'CoT?' indicates whether the model is a reasoning model. We sample 4 times for each sentence and report the averaged "COMET / COMETKiwi" scores for all models. DRT-* and DeepTrans-7B are only optimized in English (En) → Chinese (Zh) direction. **Bold** and underline denote the best and sub-optimal performance, repectively.

lines.

(3) *MT-oriented models*. We select the open-sourced DRT-7B and DRT-14B models (Wang et al., 2024a) as the baselines, which are fine-tuned by the Long CoT translation data. Besides, we also compare our model with the open-sourced DeepTrans-7B (Wang et al., 2025b), which is optimized by reinforcement learning. Further, we employ SFT to optimize the backbones with the DeepSeek-R1 generated samples of the data in ColdStart or ColdStart + RL stages, and obtain the MT- *- *ColdStart* / *FullData* models. Depending on whether involving the thinking content of DeepSeek-R1 in SFT data, we build the MT-* (w/o CoT) and MT-* (w/ CoT) models and treat the MT-*-*ColdStart* (w/ CoT) as the cold-start models of RL. Additionally, we train FeedTrans-* w/o FeedRollout models, which are optimized only with GRPO and the reward defined in Eq. 3, and use them as baselines in our RL setting. More detailed implementation of training and inference could be found in Appendix 7.1.

4.3 Main Results

We analyse the data presented in Table 1 and 2 to assess the performance of *FeedTrans-** in four translation directions and four domains in English-to-Chinese directions. Focusing on this data, we can draw several conclusions.

First, MT-oriented SFT models generally per-

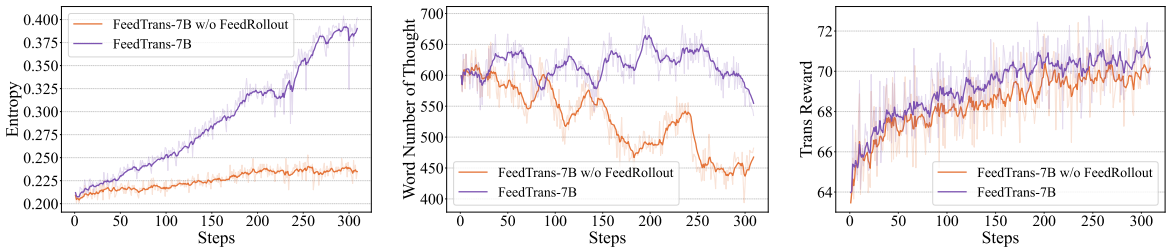
Model	CoT?	CultureMT	Tico19	MetaphorTrans	RedTrans
DeepSeek-v3	×	85.65 / 84.12	88.95 / 85.66	81.65 / 75.06	87.55 / 80.48
Backbone: Qwen2.5-7B-Instruct					
Qwen2.5-7B-Instruct	×	82.61 / 80.88	87.09 / 83.37	75.30 / 68.66	80.45 / 72.60
R1-Qwen-7B	✓	77.34 / 74.62	84.11 / 79.84	67.93 / 61.43	78.52 / 72.12
Macro-o1	✓	82.52 / 81.99	86.57 / 83.47	76.79 / 70.98	82.49 / 77.02
DRT-7B	✓	81.51 / 80.81	85.87 / 83.15	**79.93** / 72.06	81.01 / 77.14
DeepTrans-7B	✓	83.23 / 83.11	86.53 / 83.84	78.27 / 72.82	83.15 / 78.80
MT-7B-*ColdStart* (w/o CoT)	×	83.03 / 81.30	87.10 / 83.52	76.77 / 69.99	81.61 / 75.42
MT-7B-*FullData* (w/o CoT)	×	83.75 / 82.96	87.40 / 83.98	77.19 / 71.01	80.96 / 74.98
MT-7B-*ColdStart* (w/ CoT)	✓	83.53 / 82.78	86.59 / 83.17	77.42 / 71.32	84.32 / 78.48
MT-7B-*FullData* (w/ CoT)	✓	83.75 / 82.95	86.68 / 83.48	77.64 / 71.27	83.99 / 78.49
FeedTrans-7B (***Ours***)	✓	**84.13 / 83.93**	**87.69 / 84.47**	79.73 / **73.81**	**84.82 / 79.66**
w/o FeedRollout	✓	83.73 / 83.22	87.31 / 84.13	79.15 / 72.65	84.58 / 78.99
Backbone: Llama3-8B-Instruct					
Llama3-8B-Instruct	×	78.78 / 79.01	84.31 / 80.99	71.11 / 66.07	78.52 / 72.73
R1-Llama-8B	✓	79.90 / 78.87	84.12 / 80.34	72.17 / 66.14	79.72 / 73.85
MT-8B-*ColdStart* (w/o CoT)	×	80.74 / 79.28	85.66 / 81.46	72.64 / 66.21	79.54 / 73.66
MT-8B-*FullData* (w/o CoT)	×	82.01 / 80.97	85.95 / 82.11	73.97 / 66.96	79.82 / 73.82
MT-8B-*ColdStart* (w/ CoT)	✓	81.95 / 81.39	86.05 / 82.98	74.69 / 68.55	82.70 / 77.37
MT-8B-*FullData* (w/ CoT)	✓	82.92 / 82.25	86.37 / 83.10	75.83 / 69.18	82.68 / 77.31
FeedTrans-8B (***Ours***)	✓	**83.03 / 83.12**	**87.19 / 84.04**	**77.87 / 72.04**	**83.39 / 78.88**
w/o FeedRollout	✓	82.68 / 82.50	86.86 / 83.76	76.33 / 70.61	82.91 / 77.93
Backbone: Qwen2.5-14B-Instruct					
Qwen2.5-14B-Instruct	×	83.23 / 81.98	87.14 / 84.08	78.29 / 71.81	81.48 / 73.17
R1-Qwen-14B	✓	82.86 / 81.71	87.20 / 84.17	77.45 / 71.28	84.42 / 77.83
DRT-14B	✓	83.16 / 82.12	86.59 / 84.22	80.16 / 72.05	80.65 / 76.68
MT-14B-*ColdStart* (w/o CoT)	×	83.80 / 82.62	87.82 / 84.26	78.62 / 72.04	82.64 / 75.76
MT-14B-*FullData* (w/o CoT)	×	84.26 / 83.37	87.76 / 84.31	78.61 / 72.28	82.94 / 75.89
MT-14B-*ColdStart* (w/ CoT)	✓	83.97 / 83.09	87.26 / 83.86	77.95 / 71.64	84.79 / 78.71
MT-14B-*FullData* (w/ CoT)	✓	84.07 / 83.27	87.06 / 83.87	77.51 / 71.57	85.01 / 79.07
FeedTrans-14B (***Ours***)	✓	**84.45 / 84.00**	**88.14 / 85.08**	**80.53 / 74.46**	**85.71 / 80.24**
w/o FeedRollout	✓	84.22 / 83.61	87.66 / 84.61	79.46 / 73.24	85.43 / 79.69

Table 2: Translation results in four translation directions of the Flores test set. 'CoT?' indicates whether the model is a reasoning model. We sample 4 times for each sentence and report the averaged "COMET / COMETKiwi" scores for all models. **Bold** and underline denote the best and sub-optimal performance, repectively.

form better than the general-purpose models in the translation task. For instance, under the *Qwen2.5-7B-Instruct* backbone, the SFT variants (*e.g.*, *MT-7B-FullData-**) consistently outperform the backbone or *R1-Qwen-7B* across four Flores directions. Second, enabling CoT within MT-oriented SFT (*MT-*-FullData w/ CoT*) could gain over its non-reasoning counterpart on the four translation directions of Flores, but cannot achieve the consistent improvement in the other four domains of English-to-Chinese translation.

Meanwhile, the vanilla RL-based models (*i.e.*, *FeedTrans-* w/o FeedRollout*) are competitive on Flores and show clear advantages on multi-domain benchmarks (CultureMT, Tico19, MetaphorTrans, and RedTrans), which demonstrates the effectiveness and robustness of RL. However, the vanilla RL-based models cannot consistently perform better than the SFT with CoT variants (*MT-*-FullData w/ CoT*), especially on the Flores task. Differently, across backbones and scales, our *FeedTrans-7/8/14B* models achieve the best or sub-optimal results on almost all test sets, and ablating FeedRollout consistently degrades performance in four translation directions of Flores and multi-domain evaluations. For instance, in the En $\rightarrow$ De translation of Flores, the *MT-*-FullData (w/ CoT)* performs better or close to the *FeedTrans-* w/o FeedRollout*, but significantly lags behind *FeedTrans-** in terms of COMET and COMETKiwi.

Overall, the superior performance of *FeedTrans-** showcases that the FeedRollout mechanism materially contributes to better translation quality via the real-time feedback during rollout.



(a) Entropy (b) Length of Thought (c) Translation Reward (r_{trans})

Figure 4: Detailed performance changes of FeedTrans-7B during RL training. Among them, (b) and (c) display the performance of the corresponding values in the **first rollout** of FeedRollout, for fair comparison. The horizontal axis denotes the number of training steps, with 300+ steps (2 epochs) in total. The vertical axis denotes the value of the corresponding metrics.

5 Analysis

5.1 The Effect of FeedRollout

We further discuss by comparing the performance of *FeedTrans-7B w/o FeedRollout* and *FeedTrans-7B* during RL training, to provide a deeper understanding of FeedRollout. Figure 4 shows the corresponding details, which we analyze from the following aspects:

Entropy. Entropy of rollouts quantifies the stochasticity of action distributions, where a higher value reflects stronger exploratory behavior. In Figure 4(a), the entropy of the *FeedTrans-7B* rapidly increases by around 0.2 increments, while *FeedTrans-7B w/o FeedRollout* suffers a gradual rise (only 0.025 increments). Specifically, the FeedRollout mechanism leverages DeepSeek-v3 to provide quality-aware feedback according to the criterion in Figure 2. Consequently, the feedback incentivizes the policy model π to explore better translation candidates. Conversely, the minimal entropy increase in the *FeedTrans-7B w/o FeedRollout* suggests constrained exploration, indicating that relying solely on the reward score has limited capacity to guide the policy model toward improved translations. Importantly, the sustained entropy gap supports the hypothesis that the FeedRollout mechanism, through external feedback, enhances the policy model's ability to explore pathways leading to better translations.

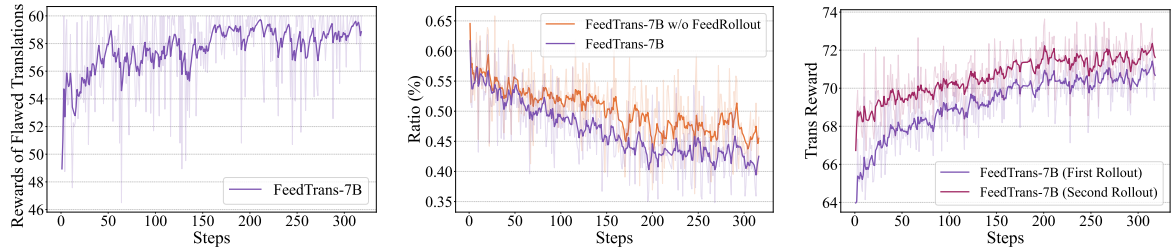
Length of Thought. In Figure 4(b), we plot the variation in the number of words in the thought process for both the *FeedTrans-7B* and *FeedTrans-7B w/o FeedRollout*. Specifically, *FeedTrans-7B w/o FeedRollout* exhibits a fluctuating yet generally declining thought length (from 600 words to 450 words), consistent with previous approaches (Feng et al., 2025b; Wang et al., 2025b). In stark contrast, *FeedTrans-7B* maintains long thought paths (over 600 words), particularly when the *FeedTrans-7B w/o FeedRollout*'s trajectory shortens abruptly. Moreover, the difference in the length variations of thought suggests that the external feedback systematically encourages the policy model to engage in deeper exploration. Besides, the sustained thought length aligns with the previously observed increase of entropy loss under FeedRollout, providing complementary evidence for the LRMs' capacity to enhance exploration during RL training. Collectively, these findings indicate that external feedback fosters a more exhaustive exploration of the translation space by the policy model, synergizing standard sampling with feedback-driven refinement to optimize candidate generation.

Translation Reward. In Figure 4(c), we present the changes in translation reward (r_{trans} in Eq.2) for both *FeedTrans-7B* and *FeedTrans-7B w/o FeedRollout* models. Specifically, *FeedTrans-7B* demonstrates a significant reward advantage after the initial training phase (*e.g.*, after 10–20 steps), maintaining consistent superior performance thereafter. In contrast, *FeedTrans-7B w/o FeedRollout* exhibits slower convergence and eventually plateaus, resulting in a reward score gap of approximately 1 to 2 points. The increase in reward suggests that the policy model could generate better translations with the expanded search space.

Takeaway-I for Effect of FeedRollout

FeedRollout could assist the policy model in expanding its search space for translation, thereby generating better translations.

Besides, we note that the behaviors of entropy loss, length of thought, and translation reward are similar between the *FeedTrans-7B* and *FeedTrans-*



(a) r_{trans} of translations in bottom 30% (b) Ratio of $r_{\text{trans}} \leq 60$ (c) r_{trans} in FeedRollout

Figure 5: Detailed performance changes of FeedRollout during RL training. Among them, (a) and (b) display the performance of the corresponding values in the **first rollout** of FeedRollout, while (c) compares the performance between the first and second rollouts. The horizontal axis denotes the number of training steps, and the vertical axis denotes the value of the corresponding metrics.

7B w/o FeedRollout during the early stages of RL, and the differences become significant in the later training steps. Based on these observations, we speculate that the effect of FeedRollout differs between the early and later training steps.

Rewards of Flawed Translations. To investigate the fine-grained effect of FeedRollout, Figure 5(a) further analyzes the variation in translation rewards for flawed translations (bottom 30% of each step) in the first rollout, as well as Figure 5(b) displays the ratio of $r_{\text{trans}} \leq 60$.

During the early training steps, depicted in Figure 5(b), the translation reward increases rapidly from approximately 50 to 58 during the initial 50 RL steps. Meanwhile, Figure 5(b) showcases that the ratio of translations with $r_{\text{trans}} \leq 60$ has the similar decline trend between *FeedTrans-7B* and *FeedTrans-7B w/o FeedRollout*. In other words, the low-quality translations are being pruned by the policy model. Building upon this phenomenon, we could plot a preliminary conjecture that FeedRollout mainly brings a systemic shift in the policy model's cognition to prune low-quality translations in the early steps, rather than immediately expanding its search space for better translations.

For the later training steps, the translation rewards of flawed translations continue to increase gradually, reaching a score of around 59. Meanwhile, the proportion of translations with $r_{\text{trans}} \leq 60$ for *FeedTrans-7B* decreases continuously to around 0.44, which is significantly lower than the *FeedTrans-7B w/o FeedRollout*'s proportion of approximately 0.5. These divergent trends following the initial phase suggest that FeedRollout transforms the policy model's early cognition to prune low-quality translations, subsequently guiding the model to focus its exploration on high-quality translations.

Reward Comparison. To further examine the fine-grained effect of FeedRollout, we present the translation rewards (r_{trans}) for both rollouts in Figure 5(c). From this figure, we observe a large discrepancy between the reward scores of the first and second rollouts during the early RL steps. In contrast, this gap becomes stable and smaller in the later steps. The significant reward deficit between the two rollouts in the early steps suggests that the initial policy model explores translations based on its own internal recognition, and that real-time feedback will assist in re-establishing its cognition. Interestingly, the gap narrows rather than expands in the later stages. This finding indicates that the core value of feedback indeed lies in the early stage of cognitive re-establishment. Moreover, the stability and smaller gap in the later stages precisely suggest that the new cognition guided by feedback is persistent, enabling the policy model to continue exploring better translations without becoming overly dependent on external feedback.

Takeaway-II for Effect of FeedRollout

The impact of FeedRollout varies between the initial and later stages of training. Initially, FeedRollout reshapes the policy model's understanding to eliminate low-quality translations. As training progresses, FeedRollout directs the model's exploration towards high-quality translations.

5.2 Comparison with Large-Scale Models

We further analyze the performance of *FeedTrans-7B/14B* in comparison with larger models (*e.g.*, *QwQ-32B*, *GPT-4o*, *o1-preview*, and *DeepSeek-R1*). To reduce the cost of APIs, we randomly select 400

Model	Param.	Flores200	CultureMT	Tico19	MetaphorTrans	RedTrans
GPT-4o	-	87.45 / 84.15	84.18 / 83.14	85.66 / 81.36	70.70 / 66.06	83.55 / 78.69
o1-preview	-	**89.29 / 86.18**	85.80 / **84.89**	**88.79 / 84.47**	**80.77** / 74.41	**86.52** / 81.61
DeepSeek-R1	671B	89.23 / **86.18**	**86.13** / 84.63	88.10 / 83.95	79.04 / 72.81	85.88 / 81.08
Qwen2.5-32B-Instruct	32B	87.22 / 83.90	84.31 / 80.47	87.49 / 83.53	78.53 / 72.60	84.91 / 79.88
R1-Qwen-32B	32B	88.49 / 85.31	84.11 / 80.47	87.67 / 83.96	78.97 / 72.52	85.44 / 80.77
QwQ-32B	32B	88.75 / 85.92	85.19 / 84.62	87.98 / 83.89	79.89 / 73.52	85.71 / 81.35
FeedTrans-7B	7B	88.62 / 85.93	84.78 / 84.42	87.41 / 83.69	79.69 / 73.62	85.53 / 81.90
FeedTrans-14B	14B	88.93 / 86.04	85.30 / 84.54	87.40 / 83.88	80.57 / **74.58**	85.95 / **82.15**

Table 3: Comparison results in English-to-Chinese translation of FeedTrans-7B/14B and strong baseline models (commercial LLMs or >30B parameters). In this table, we report the scores by evaluating models on the randomly selected 400 samples for each test set to reduce the cost of APIs.

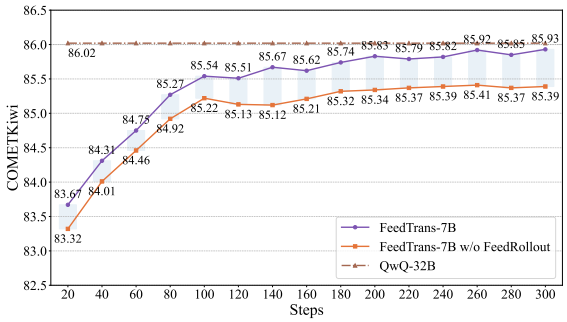


Figure 6: COMETKiwi scores of FeedTrans-7B and FeedTrans-7B *w/o FeedRollout* among training steps on the Flores200.

Model	Performance	GPU hours
MT-7B-*FullData* (w/o CoT)	83.45 / 79.56	3.3 + 25.6
MT-Patcher-7B (w/o CoT)	84.12 / 80.34	3.3 + 58.3
MT-7B-*FullData* (w/ CoT)	84.04 / 80.23	3.3 + 25.6
MT-Patcher-7B (w/ CoT)	84.75 / 80.87	4.5 + 58.3
FeedTrans-7B	**85.38 / 81.73**	113.4 + 76.8
w/o FeedRollout	84.64 / 81.03	86.2 + 62.3

Table 4: Exploration of different FeedBack Mode and the GPU cost. Performance is the average scores (COMET/COMETKiwi) among five test sets in En → Zh translation (Flores, CultureMT, Tico19, MetaphorTrans, and RedTrans). "a + b" in GPU hours denotes the "trained model time" (all time of trained model) + "auxiliary model time" (all time of other auxiliary model, *e.g.*, reward model).

samples in each test set and list the results in Table 3.

Compared to the larger models, *FeedTrans-7B/14B* could approach or even surpass the performance of *o1-preview* and match that of *QwQ-32B* in our scenarios. For example, on the WMT23 dataset, *FeedTrans-7B* achieves a score of 85.06 / 81.67, comparable to *QwQ-32B*'s 84.95 / 81.36 and surpassing some metrics of larger models like *DeepSeek-R1*. On the Flores200 dataset, *FeedTrans-7B* attains a score of 88.62 / 85.93, closely matching *QwQ-32B* and approaching *o1-preview*'s leading score of 89.29/86.18. *FeedTrans-14B* further enhances these results, scoring 80.57 / 74.58 on the MetaphorTrans dataset, which surpasses *QwQ-32B* and approaches *o1-preview*. These findings indicate that FeedTrans models, despite their smaller size, achieve translation quality close to significantly larger models, highlighting the effectiveness and scalability of our *FeedTrans-7B/14B* model.

5.3 Performance among Steps

In this section, we present the COMETKiwi scores of *FeedTrans-7B* and *FeedTrans-7B w/o FeedRollout* across training steps, as shown in Figure 6.

During the early stages of training, *FeedTrans-7B* demonstrates a statistically significant and stable advantage over the model without FeedRollout, maintaining a consistent lead of approximately 0.36 in COMETKiwi scores. Furthermore, the performance divergence intensifies in the later training steps, with the advantage of *FeedTrans-7B* expanding by 46% to a final gap of 0.54 COMETKiwi points. Moreover, the trend is consistent with the findings in §5.1, providing additional evidence of the impact of FeedRollout at different stages of training. Notably, *FeedTrans-7B* achieves performance similar to that of QwQ-32B (85.93 vs. 86.02 COMETKiwi scores), despite having significantly fewer parameters (7B vs. 32B). Overall, the superior performance of *FeedTrans-7B* underscores the effectiveness of FeedRollout in enhancing the translation capabilities of the policy model.

5.4 Feedback Usage

Further, we analyze the usage of feedback from other models and GPU hours during training and list the results in Table 4. In detail, we compare four

Reward Model	FeedTrans-7B w/o FeedRollout	FeedTrans-7B
DeepSeek-v3	84.62 / 80.88	85.04 / 81.55
+ COMETKiwi	**84.89 / 81.09**	**85.38 / 81.73**
DeepSeek-R1	84.64 / 81.03	84.98 / 81.60
QwQ-32B	84.43 / 80.61	84.78 / 81.32

Table 5: Averaged scores of RL models (based on *MT-7B-FullData w/ CoT*) with different reward models among five test sets in En → Zh translation.

ways of feedback: direct SFT (*MT-7B-FullData-**), offline feedback & SFT (*MT-Patcher-7B-**)[1] (Li et al., 2024), online judge (*FeedTrans-7B w/o FeedRollout*), and online feedback (*FeedTrans-7B*).

As shown in Table 4, in terms of performance, the usage of feedback always performs better than direct SFT, no matter the offline or online feedback. Besides, although the offline feedback could significantly improve the performance compared to *MT-7B-FullData-**, it still lags behind the online feedback. This performance gap further demonstrates the effectiveness of our proposed FeedRollout mechanism. For the GPU hours cost, the primary expense for the SFT-based models (*MT-7B-FullData-* and *MT-7B-Patcher-*) comes from the auxiliary model (*e.g.*, *DeepSeek-R1*) used to generate or consolidate the SFT data. Compared with SFT, the RL-based models consume more GPU hours because they rely on rollouts and the corresponding interactions with the reward model. Moreover, due to the refinement and extra interaction with the reward model, the online feedback requires more GPU hours to achieve the best performance.

5.5 Reward Model Selection

In this section, we investigate the impact of the reward model on the final performance and compare the basic selection (*i.e.*, *DeepSeek-v3*) with three strategies, *i.e. DeepSeek-v3 + COMETKiwi* (averaged scores), *DeepSeek-R1*, and *QwQ-32B*. The results are listed in Table 5.

Table 5 shows that *FeedTrans-7B* consistently outperforms *FeedTrans-7B w/o FeedRollout* under all reward model strategies, indicating that our proposed FeedRollout is not limited to the *DeepSeek-v3* and performs well under different reward models. In addition, *MT-7B-FullData (w/ CoT)* achieves 84.04 / 80.23 average scores in Table 4, which lags behind *FeedTrans-7B* using

[1]We describe the implementation in Appendix 7.1.

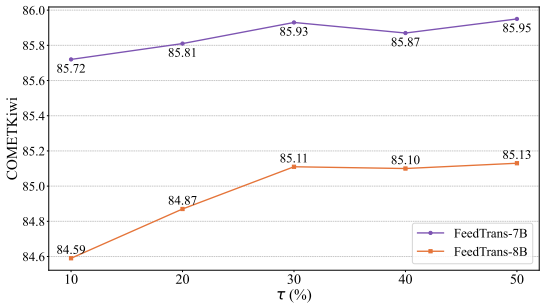


Figure 7: Performance of the different selected ratios by τ of "Flawed" translation on the English-to-Chinese test set of Flores200.

DeepSeek-R1 as the reward model (84.98 / 81.60). This suggests that, relative to SFT improvements obtained mainly via distilled data, RL with FeedRollout can leverage real-time feedback during training more effectively, leading to stronger translation performance.

5.6 Hyperparameter Experiments

As introduced in § 3.2, FeedRollout introduces the hyperparameter τ to select the "Flawed" translations. For both backbones (*Qwen2.5-7B-Instruct* and *Llama3-8B-Instruct*), we vary τ from 10% to 50%, and evaluate on the English-to-Chinese translation test set of Flores200. The results in Figure 7 indicate that τ=30% is a strong choice under our training setup. Specifically, the smaller τ values constrain FeedRollout's potential by providing less feedback to support reflective reasoning. Conversely, increasing τ beyond 30% does not yield substantial additional gains. Empirically, with $\tau \geq$ 30%, the policy model produces improved translations for roughly 50% of the identified flawed cases in each rollout. Therefore, we adopt τ=30% as the default setting to train our translation models, enabling them to learn self-assessment from *DeepSeek-v3* and achieve strong performance.

5.7 Effect of Secondary Decoding with Feedback

Since we apply secondary rollout with feedback during RL training, in this section, we analyze the effect of secondary decoding with feedback during inference. We explore three models: *MT-7B-Cold Start (w/ CoT)*, which denotes the model SFT with the long-thought MT data; our *FeedTrans-7B* and *FeedTrans-7B w/o FeedRollout*.

As shown in Table 6, we observe that the improvement of *MT-7B-Cold Start (w/ CoT) + feed-*

Metrics	MT-7B-*ColdStart* (w/ CoT)	FeedTrans-7B w/o FeedRollout	FeedTrans-7B
Vanilla Infer	86.82 / 82.97	88.34 / 85.42	88.81 / 85.89
+ SelfAssess	+0.31 / +0.77	-0.11 / -0.18	+0.12 / +0.02
+ FeedBack	+0.66 / +2.96	+0.13 / +0.26	+0.15 / +0.03
Number of Reflection	0.87	1.33	2.36

Table 6: Analysis of the reflective behavior in three models on the Flores200 task. "Vanilla Infer" reports the COMET/COMETKiwi score by the vanilla inference. "SelfAssess" denotes the improvement of "Vanilla Infer' with the self-assessment. "FeedBack" denotes the improvement of "Vanilla Infer" with the external feedback from DeepSeek-v3, where the operation is the same as the flawed translation in FeedRollout for each case in the testset. "Number of Reflection" presents the average number of reflection behavior in "Vanilla Infer", detected by keywords like "Wait".

back indicates that incorporating external feedback from DeepSeek-v3 during inference enhances the model's translation quality. In contrast, for our *FeedTrans-7B*, external feedback yields minimal gains. Additionally, we investigate model performance when using self-generated feedback, *i.e.*, "vanilla infer + self-assess". In this setting, *FeedTrans-7B* achieves comparable results under both "vanilla infer" and "vanilla infer + feedback," whereas *FeedTrans-7B w/o FeedRollout* fails to benefit from self-feedback and shows no further improvement. We further calculate the "Number of Reflections" for different models. Notably, *FeedTrans-7B* exhibits a higher average number of self-reflections compared to *MT-7B-Cold Start (w/ CoT)*. This phenomenon suggests that *FeedTrans-7B* effectively learns to self-assess and refine translations with the assistance of FeedRollout during training, without requiring human intervention in secondary decoding during inference.

5.8 Human Evaluation

We conduct human evaluation to further assess the performance of *Qwen2.5-7B-V3*, *FeedTrans-7B*, and *FeedTrans-7B w/o FeedRollout*. We randomly select 200 samples from Flores200 and CultureMT, and employ five human evaluators with high levels of fluency in English and Chinese to assess the generated translations. The human evaluation focuses on three key aspects: fluency, semantic accuracy, and terminology selection. Following Kiritchenko and Mohammad (2017), evaluators assess the quality of each aspect and rank the candidate translations for each source input, with rank 1 indicating the best one.

Model	Fluency	Semantic	Terminology
Flores200			
Qwen2.5-7B-*V3*	2.57	2.24	1.99
FeedTrans-7B	**1.51**	**1.79**	**1.87**
w/o FeedRollout	1.92	1.97	2.09
CultureMT			
Qwen2.5-7B-*V3*	2.31	2.15	2.01
FeedTrans-7B	**1.61**	**1.81**	**1.82**
w/o FeedRollout	2.08	2.04	2.17

Table 7: Human evaluation results in terms of fluency, semantic accuracy, and terminology selection. Lower values indicate better performance.

Table 7 reports the average rank of each model on both test sets. The results indicate that *FeedTrans-7B* outperforms *Qwen2.5-7B-V3* and *FeedTrans-7B w/o FeedRollout* in both fluency and semantic accuracy across both Flores200 and CultureMT. Noteably, *FeedTrans-7B w/o FeedRollout*, which only applies GRPO with final result scoring, shows markedly poorer terminology selection than *FeedTrans-7B*. This suggests that the FeedRollout mechanism, which provides in-rollout feedback to generated translations, is crucial for improving terminology selection during RL. In contrast, *Qwen2.5-7B-V3*, trained solely via SFT with data distilled from DeepSeek-v3, performs better than *FeedTrans-7B w/o FeedRollout* on terminology selection but worse on fluency and semantic accuracy. Overall, these comparisons demonstrate that incorporating FeedRollout during RL training not only enhances translation fluency and semantic accuracy but also improves terminology selection.

5.9 Case Study

Table 8 presents a case study to evaluate the practical translation performance of *Qwen2.5-7B-V3*, *FeedTrans-7B w/o FeedRollout*, and *FeedTrans-7B*. Besides, we list the translation results of *DeepTrans-7B* and *QwQ-32B* for comparison.

In Case 1, the sentence contains specialized terms such as "Klansman" that require precise translation. While *Qwen2.5-7B-V3* captures the general meaning, it mistranslates "Klansman" and "role", leading to semantic inaccuracies. Among the RL-based models, only *FeedTrans-7B* correctly translates "Klansman" as "三K党", demonstrating stronger domain knowledge. Other models produce inaccurate or imprecise translations, reflecting limitations in handling proper nouns and specific terms. Additionally, *FeedTrans-7B w/o FeedRollout* ex-

Source Sentence	Groud Truth	Qwen2.5-7B-V3	FeedTrans-7B w/o FeedRollout	FeedTrans-7B	DeepTrans-7B	QwQ-32B
			Case 1			
He's played the part of the head Klansman for years, but he said the role never gets any easier.	多年来，他一直扮演着三K党头目的角色，但他说这个角色从不轻松。	他已扮演多年头号吉姆·克劳斯人的角色，但表示这一行永远不会轻松。	他已多年扮演白人民族主义头目这一角色，但他表示，无论时间如何推移，这个角色始终难以轻松应对.	他多年来一直在扮演三K党头目的角色，但他说这个角色永远都不会变得轻松。	他扮演这个头号凯尔森人已经多年了，但坦言，这个角色始终没有变得轻松起来。	他多年来一直饰演克兰组织首领的角色，但他表示，这个角色从未变得轻松一些。
		(He has played the leading role of "Jim Klausen" for many years, but he says this profession will never be easy.)	*(He has played the role of the leader of the "white nationalist movement" for many years, but he said that no matter how time goes by, this role is always difficult to handle easily.)*	*(He has been playing the role of the leader of the "Klansman" for many years, but he said that this role will never become easy.)*	*(He has been playing this top-level Kelsenian for many years, but he admits that this role has never become any easier.)*	*(For many years, he has been playing the role of the leader of the Kran Organization, but he said that this role has never become any easier.)*
			Case 2			
Dynamic on vault.	跳马时充满活力。	动态式保险箱。	金库动态。	在跳马上展现出动态活力。	仓储区域充满活力。	在跳马项目中的动态表现
		(Dynamic-style Safe Box.)	*(Coffer Dynamics.)*	*(Showing dynamic vitality on vault.)*	*(The warehouse area is full of vitality.)*	*(The dynamic performance in the vault event.)*
		N/A	..., Here, the term "vault" could refer to a bank vault, meaning of "金库", ...	..., "Vault" usually refers to "储藏室" or "金库", but "Dynamic on vault" might be describing an action or state that is both dynamic and energetic. For example, in gymnastics, the "Vault" event ("跳马")...	..., the word "vault" of "on vault" refers to a "仓库" or "储藏室", ...	"Dynamic" can mean something that's energetic, active, or changing,..."Vault" here is likely referring to the vaulting event in gymnastics,..., where "跳马" is the term for vault in gymnastics.

Table 8: Case Studies, Purple and Green denote the error and correct translation, respectively. We provide the additional "*(English explanation)*" for each Chinese translation. Besides, Gray words present the corresponding thought content of each model.

hibits an over-translation error (*i.e.*, "无论时间如何推移"/"no matter how time goes by"), rather than the terminology selection error.

Case 2 poses a challenge due to the lexical ambiguity of "vault", which can mean either a "bank vault" or a gymnastics event. Most models fail to resolve this ambiguity without sufficient context, translating it as "金库" or "仓储区域", which are incorrect in this context. In contrast, *FeedTrans-7B* and *QwQ-32B* successfully interpret "vault" as the gymnastics event "跳马" ("vaulting horse"). Notably, *FeedTrans-7B* demonstrates deeper reasoning by explicitly recognizing and successfully disambiguating the term as "跳马". This demonstrates that *FeedTrans-7B*, trained with FeedRollout, exhibits advanced semantic understanding and improved ambiguity resolution through reflective reasoning.

Overall, *FeedTrans-7B* produces translations with higher semantic completeness and more natural fluency. These findings suggest that our FeedRollout mechanism could effectively enhance the model's ability to understand and reflect on complex linguistic elements, thereby improving overall translation quality.

6 Conclusion

In this paper, we investigate how translation evaluation feedback assists large reasoning models in achieving better translation performance, to overcome the lack of translation guidance in MT-oriented RL. Specifically, we propose a novel FeedRollout mechanism to incorporate feedback as guidance, to engage the policy model in thoughtful reasoning towards higher-quality translations. Experimental results on six translation tasks demonstrate the effectiveness of FeedTrans-7/8/14B with the assistance of FeedRollout. Besides, we conduct a detailed analysis to demonstrate that FeedRollout expands the policy models' search space and enhances their translation quality. Furthermore, we explore and summarize the different effects of FeedRollout in the early and later RL steps.

Batch Size	MT-7B-*ColdStart* (w/o CoT)	MT-7B-*FullData* (w/o CoT)	MT-7B-*ColdStart* (w/ CoT)	MT-7B-*FullData* (w/ CoT)
		learning rate = 1e-5		
8	83.24 / 78.85	83.45 / 79.56	**83.94** / 80.09	84.04 / 80.23
32	**83.32** / 78.71	83.53 / **79.65**	83.89 / 79.98	84.06 / 80.22
64	83.18 / 78.44	**83.64** / 79.61	83.84 / 80.01	84.01 / 80.24
128	83.23 / **79.01**	83.59 / 79.57	83.92 / **80.10**	84.08 / 80.20
256	83.11 / 78.92	83.61 / 79.59	83.86 / 79.41	**84.10** / **80.26**
		learning rate = 1e-6		
8	83.25 / 79.09	83.56 / 79.41	**83.49** / **79.64**	**84.08** / 80.17
32	83.22 / **79.13**	**83.66** / 79.46	82.45 / 79.55	84.04 / 80.20
64	**83.33** / 79.05	83.65 / **79.49**	83.41 / 79.47	84.02 / **80.22**
128	83.13 / 78.93	83.61 / 79.45	83.15 / 79.39	83.91 / 80.18
256	82.82 / 78.56	83.57 / 79.33	82.19 / 78.21	83.58 / 79.90

Table 9: Averaged score of five test sets in En → Zh translation (Flores, CultureMT, Tico19, MetaphorTrans, and RedTrans). 'CoT?' indicates whether the model is a reasoning model. We sample 4 times for each sentence and report the averaged "COMET / COMETKiwi" scores for all models. **Bold** denotes the best performance.

7 Appendix

7.1 Implementation Details

SFT. We use Llama-Factory (Yang et al., 2024) to perform SFT, producing the distilled data (ColdStart and RL stages) from DeepSeek-R1. DeepSpeed ZeRO-3 (Rasley et al., 2020) is also set to save memory. During training, we firstly prepare the ColdStart models (MT-*-*ColdStart*) by SFT the backbones with the ColdStart data. Consequently, we conduct the distilled variants (MT-*-*FullData*) by SFT the ColdStart models with the RL stages data, for the comparison with the RL models. For the MT-Patcher models (MT-Patcher-7B-* in §5.4), we leverage the ColdStart models to sample translations and employ DeepSeek-v3-awq to provide feedback and refinement, then SFT ColdStart models with the merged data, following Li et al. (2024). For training details, we set the batch size to 8 and the learning rate to 1e-5[2], then train all models in 2 epochs.

RL Training. We use the GRPO RL algorithm implemented by VeRL (Sheng et al., 2024), based on MT-*-*ColdStart* (w/ CoT). During training, we configure a batch size of 64 and set N of rollout to 8, and each rollout result is evaluated by the deployed DeepSeek-v3 (awq quantization). We set the learning rate to 1e-6, the sampling temperature to 0.6, and the maximum generation length to 4096 tokens. Then, we train the model 2 epochs on 8×NVIDIA H20, and the DeepSeek-v3 is deployed on 8×NVIDIA H20. In practice, we treat the bottom 30% (*i.e.*, the bottom τ-%) of translations within each batch as flawed and refine them using FeedRollout. We refer to the resulting models as FeedTrans-*. Additionally, we train FeedTrans-* w/o FeedRollout models, which are optimized only with GRPO and the reward defined in Eq. 3, and use them as baselines in our RL setting.

Inference. To accelerate the inference, we employ vLLM (Kwon et al., 2023) to evaluate all models on 8×NVIDIA H20, setting the temperature to 0.6, top p to 0.95, repetition penalty to 1.05, and the maximum number of generated tokens to 4096.

7.2 Hyperparameter of SFT

To explore the impact of hyperparameters during training on the performance of SFT models, we conduct these experiments by searching the batch size within [8, 32, 64, 128, 256] and the learning rate within [1e-5, 1e-6], based on *Qwen2.5-7B-Instruct*.

As shown in Table 9, expensive experimental results display that, in our scenario, the values of batch size and learning rate have an insignificant impact on the *MT-7B-FullData-** with less than 0.1 in delta. Hence, to simplify the setting in our experiments, we set the batch size to 8 and the learning rate to 1e-5 for all SFT models.

[2]We explore these settings in §7.2.

Acknowledgments

The research work described in this paper has been supported by the National Nature Science Foundation of China (No. 62476023, 62376019, and 62406020), Beijing Natural Science Foundation (No. L252034), and CHN RAILWAY GRANT (No. L2023G013). The authors would like to extend their sincere gratitude to action editor Barry Haddow and all anonymous TACL reviewers for their valuable comments and suggestions to improve this paper.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, et al. 2020. Tico-19: the translation initiative for covid-19. *arXiv preprint arXiv:2007.01788*.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Dual-reflect: Enhancing large language models for reflective translation through dual learning feedback mechanisms. *arXiv preprint arXiv:2406.07232*.

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

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

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

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025a. Retool: Reinforcement learning for strategic tool use in llms. *arXiv preprint arXiv:2504.11536*.

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. 2025. Deeprag: Thinking to retrieval step by step for large language models. *arXiv preprint arXiv:2502.01142*.

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

Hongcheng Guo, Fei Zhao, Shaosheng Cao, Xinze Lyu, Ziyan Liu, Yue Wang, Boyang Wang, Zhoujun Li, Chonggang Lu, Zhe Xu, and Yao Hu. 2025b. Redefining machine translation on social network services with large language models.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model.

Hui Huang, Yancheng He, Hongli Zhou, Rui Zhang, Wei Liu, Weixun Wang, Wenbo Su, Bo Zheng, and Jiaheng Liu. 2025. Think-j: Learning to think for generative llm-as-a-judge. *arXiv preprint arXiv:2505.14268*.

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

Svetlana Kiritchenko and Saif M Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. *arXiv preprint arXiv:1712.01765*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jiahuan Li, Shanbo Cheng, Shujian Huang, and Jiajun Chen. 2024. Mt-patcher: Selective and extendable knowledge distillation from large language models for machine translation. 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 6445–6459.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025a. Torl: Scaling tool-integrated rl. *arXiv preprint arXiv:2503.23383*.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025b. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Yunlong Liang, Fandong Meng, Jiaan Wang, and Jie Zhou. 2025. Slangdit: Benchmarking llms in interpretative slang translation. *arXiv preprint arXiv:2505.14181*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. 2025b. There may not be aha moment in r1-zero-like training — a pilot study. `https://oatllm.notion.site/oat-zero`. Notion Blog.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. Notion Blog.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. Toolrl: Reward is all tool learning needs. *arXiv preprint arXiv:2504.13958*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. *arXiv preprint arXiv:2305.14878*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

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

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *arXiv preprint arXiv:2209.06243*.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. 2025a. Otc: Optimal tool calls via reinforcement learning. *arXiv preprint arXiv:2504.14870*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2024a. Drt-o1: Optimized deep reasoning translation via long chain-of-thought. *arXiv e-prints*, pages arXiv–2412.

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

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

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2024b. Delta: An online document-level translation agent based on multi-level memory. *arXiv preprint arXiv:2410.08143*.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*.

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

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

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2023. Benchmarking machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*.

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

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*.

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.