TAT-R1: Terminology-Aware Translation with Reinforcement Learning and Word Alignment

Anonymous ACL submission

Abstract

Recently, deep reasoning large language models (LLMs) like DeepSeek-R1 have made significant progress in tasks such as mathematics and coding. Inspired by this, several studies have employed reinforcement learning (RL) to enhance models' deep reasoning capabilities and improve machine translation (MT) quality. However, the terminology translation, an essential task in MT, remains unexplored in deep reasoning LLMs. In this paper, we propose **TAT-R1**, a terminology-aware translation model trained with reinforcement learning and word alignment. Specifically, we first extract the keyword translation pairs using a word alignment model. Then we carefully design three types of rule-based alignment rewards with the extracted alignment relationships. With those alignment rewards, the RL-trained translation model can learn to focus on the accurate translation of key information, including terminology in the source text. Experimental results show the effectiveness of TAT-R1. Our model significantly improves terminology translation accuracy compared to the baseline models while maintaining comparable performance on general translation tasks. In addition, we conduct detailed ablation studies of the DeepSeek-R1-like training paradigm for machine translation and reveal several key findings. The code, data, and models will be publicly released[1].

1 Introduction

Terminology translation is an essential task in machine translation, and its accuracy significantly impacts the translation quality of specialized domain texts. Many researchers have conducted extensive studies on terminology translation, proposing various methodologies. Kim et al. (2024) detect terms, constructs a terminology database, and provides term information via retrieval-augmented generation (RAG) before model translation. Moslem et al. (2023) synthesize bilingual data containing terms, fine-tunes the model, and applies post-processing to correct terminology after translation. DragFT (Zheng et al., 2024) employ few-shot examples to enhance translation performance in specialized domains. Bogoychev and Chen (2023) improve term translation by constraining incorrect terminology during decoding. These methods generally rely on relatively accurate terminology extraction to either (1) construct training data for supervised fine-tuning or (2) incorporate relevant terminological information during the inference phase.

Recent advances have demonstrated promising progress in leveraging reinforcement learning (RL) to stimulate models' deep reasoning capabilities, exemplified by DeepSeek-R1 (DeepSeek-AI, 2025). These developments have further validated that the enhanced model abilities acquired through RL exhibit strong generalization performance. Inspired by DeepSeek-R1 (DeepSeek-AI, 2025), some studies have tried to use reinforcement learning to stimulate the model's deep reasoning capabilities and improve translation quality. R1-T1 (He et al., 2025) synthesize training data with reasoning processes for translation, first applying SFT and then conducting reinforcement training using COMET (Rei et al., 2020) as the reward. Similar to DeepSeek-R1-Zero, MT-R1-Zero (Feng et al., 2025) directly perform reinforcement training on a pretrained model, employing BLEU (Papineni et al., 2002) and COMETKiwi (Rei et al., 2022) as rewards. DeepTrans (Wang et al., 2025) directly uses DeepSeek-V3 (DeepSeek-AI, 2025) scoring as the reward, enhancing the model's performance in literary translation through reinforcement learning. To the best of our knowledge, no existing research has explored the integration of reinforcement learning and deep reasoning for terminology translation tasks.

[1] https://github.com/MrHungry/TAT-R1

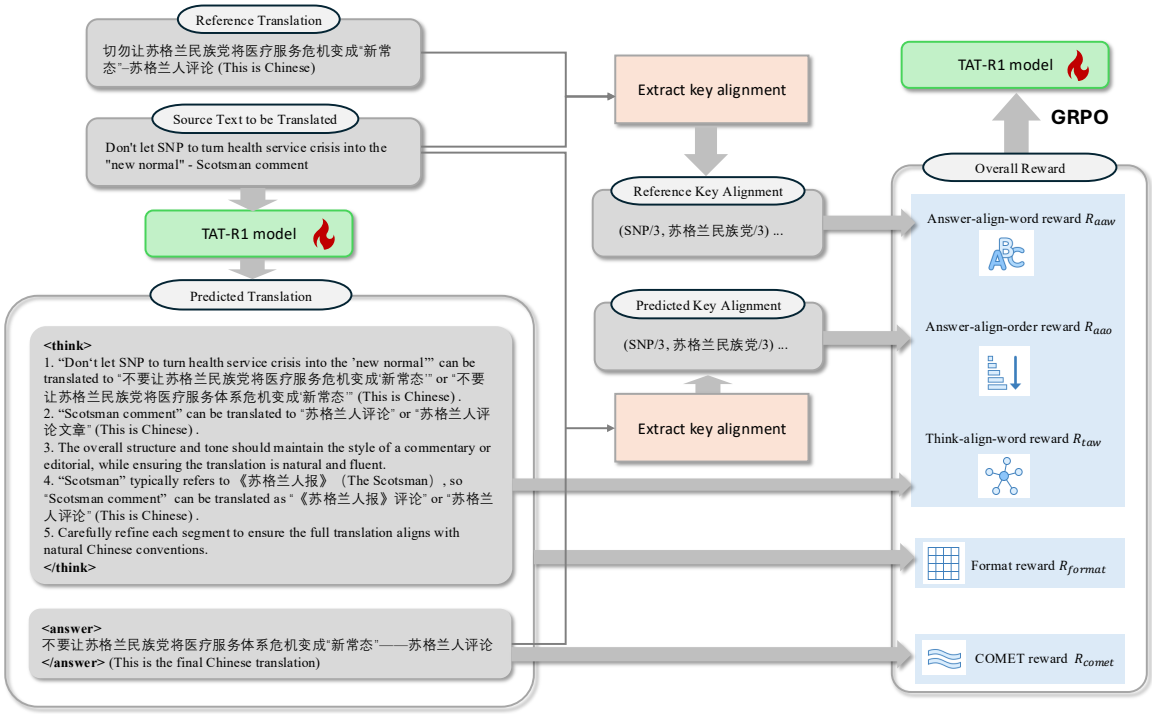

Figure 1: The overview of TAT-R1 training with RL and word alignment.

In this paper, we propose TAT-R1, a terminology-aware translation model trained with reinforcement learning and word alignment. First, using word alignment techniques, we design effective reinforcement learning reward signals for terminology translation tasks. Word alignment involves analyzing parallel bilingual corpora to determine translational equivalence between words across languages. By leveraging word alignment techniques, we can effectively extract domain-specific key terms from parallel training corpora, thereby substantially mitigating the challenge of scarce terminology-labeled training data. Then, we directly train our model using RL, and extensive experimental results demonstrate the effectiveness of our proposed method. Our main contributions are as follows:

- We propose TAT-R1, the first terminology-aware translation model trained with RL and word alignment rewards. Leveraging word alignment, we design three simple yet effective reward functions for terminology translation model training.
- Experimental results demonstrate the effectiveness of TAT-R1. TAT-R1 significantly improves terminology translation accuracy compared to the baseline while maintaining comparable performance on general translation tasks. Moreover, we do not need any terminology detection during inference.
- We conduct detailed ablation studies of the DeepSeek-R1-like training paradigm for machine translation, and reveal several key findings, including the generalization capability of RL, the different impacts of various rewards, and the effectiveness of the reasoning process.

2 Methods

In this section, we present the reward mechanisms and reinforcement learning algorithm employed in our proposed TAT-R1 model.

2.1 Design of Rewards

As shown in Figure 1, the rewards we use in our RL training have three parts: format reward, comet reward, and word alignment reward.

Format Reward. As shown below, we employed a template similar to that used in DeepSeek-R1, requiring the model to output its reasoning process within `<think></think>` tags and the translation results within `<answer></answer>` tags. Here, target_language specifies the target language, while source_text denotes the input text. To prevent the model from generating non-translation content in the `<answer></answer>` section, we explicitly in-

cluded the instruction "without additional explanations" in the User prompt.

Template for TAT-R1

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 the answer is enclosed within `<answer></answer>` tags, i.e., `<think>` reasoning process here `</think><answer>` answer here `</answer>`.

User:
Translate the following text into {target_language} without additional explanations:
{source_text}

Assistant:

We employ regular expressions to verify whether the model's output conforms to the format specified in the template. If compliant, the format reward is set to 1; otherwise, it is 0. Specifically:

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

COMET Reward. COMET is a widely-used evaluation metric in machine translation that assesses translation quality at the semantic level. The effectiveness of COMET-based rewards has been validated in papers (He et al., 2025) and (Feng et al., 2025). In this work, we incorporate COMET-22 as one component of our reward functions. To maintain training stability, we adopt a similar approach to that used in (He et al., 2025), specifically:

$$R_{comet} = round(comet, 2) \quad (2)$$

Word Alignment Reward. Semantic evaluation metrics like COMET primarily assess the overall translation quality of a model but often fail to accurately capture the translation accuracy of localized information, such as technical terms. BLEU, an n-gram-based metric, mainly measures the n-gram overlap between reference translations and model outputs. However, for translation tasks, BLEU imposes overly strict requirements. Unlike mathematics or code, where there is a single correct answer, translation permits multiple valid renditions. Enforcing strict n-gram matching between model outputs and references—as BLEU does—may not always be reasonable and could even introduce negative semantic effects, as we later verify in our experiments. For instance, a single Chinese sentence may have multiple valid English translations with varying syntactic structures.

Nevertheless, key elements like terminology often demand precise translation. By incorporating reward signals that specifically evaluate the accuracy of such critical terms, we can enhance their translation fidelity without compromising overall semantic quality.

In machine translation, word alignment is a critical task that aims to automatically establish correspondences between words in source and target language sentences. This task involves analyzing parallel bilingual corpora to determine translational equivalence between words across languages. In this work, we leverage word alignment to design three distinct reward mechanisms for improving translation quality. Next, we present the detailed computation process for the three word-alignment-based reward mechanisms.

First, we can use word alignment models to identify word-level alignment information between the source text, reference text, and translated text.

Assume the tokenized sequence of the source text is S and s_i is the i-th word in S. Therefore, the tokenized sequence of the source text can be represented as:

$$S = [s_1, s_2, s_3, ..., s_i, ..., s_N] \quad (3)$$

where N represents the length of source sequence. Similarly, the tokenized sequence of the corresponding reference translation and predicted translation can be represented as:

$$R = [r_1, r_2, r_3, ..., r_j, ..., r_M] \quad (4)$$

$$P = [p_1, p_2, p_3, ..., p_k, ..., p_K] \quad (5)$$

where M represents the length of reference tokens and K represents the length of predicted tokens. Given the tokenized sequences of the source text, reference translation, and predicted translation we can input them into the word alignment model to obtain token-level alignment relationships:

$$A^{ref} = Align(S, R) = [..., A_{ij}^{ref}, ...] \quad (6)$$

$$A^{pre} = Align(S, P) = [..., A^{pre}_{ik}, ...] \quad (7)$$

where A^{ref} represents the alignments between source and reference tokens, and A^{pre} represents the alignments between source and predicted tokens. A^{ref}_{ij} represents the i-th token in source tokens and the j-th token in reference tokens are aligned. A^{pre}_{ik} represents the i-th token in source tokens and the k-th token in predicted tokens are aligned, which can be expressed in formulas respectively as $A^{ref}_{ij} = (s_i/i, r_j/j)$ and $A^{pre}_{ik} = (s_i/i, p_k/k)$.

Next, we perform Named Entity Recognition (NER) on the source tokens, retaining nouns as key elements requiring alignment. This is because noun translations typically exhibit less variability compared to sentence structures or conjunctions, where multiple valid translations often exist. So, after this step, the key alignments are a subset of original alignments, and the aligned source tokens are all nouns. We denote the set of key alignments as A^{ref_key} and A^{pre_key}.

Last, we can calculate the alignment reward with A^{ref_key} and A^{pre_key}. As shown in Figure 1, we design three types of alignment rewards, namely answer-align-word reward R_{aaw}, answer-align-order reward R_{aao}, and think-align-word reward R_{taw}.

Answer-align-word reward reflects the word overlap ratio between the reference key alignment and the predicted key alignment. This reward encourages the model to translate key information in its output accurately and can be denoted as:

$$R_{aaw} = \frac{len(A^{ref\_key} \cap A^{pre\_key})}{len(S) + len(P)} \quad (8)$$

we include the length of the model's output token sequence in the denominator to prevent the model from generating excessively long outputs to hack this reward.

The answer-align-order reward is a reward that reflects the order overlap ratio between the reference key alignment and the predicted key alignment. This reward encourages the model to follow the order of key information as it appears in the reference translation and can be denoted as:

$$R_{aao} = \frac{len(OD(A^{ref\_key}) \cap (OD(A^{pre\_key}))}{len(OD(A^{ref\_key}))} \quad (9)$$

where $OD(X)$ means getting the order pairs of sequence X. For example, if $X = [a, b, c]$, then $OD(x) = \{ab, ac, bc\}$.

Similar to the answer-align-word reward, the think-align-word reward is a reward that reflects the word overlap ratio between the reference key alignment and the text in tags <think> and </think>. This reward encourages the model to consider how to translate key information before outputting the final answer and can be denoted as:

$$R_{taw} = \frac{num\ of\ A^{ref\_key}\ hit\ in\ think}{len(A^{ref\_key})} \quad (10)$$

where $num\ of\ A^{ref\_key}\ hit\ in\ think$ means the number of appearances of aligned word pairs in text between <think> and </think> tags. For example, if one item of A^{ref_key} is $A^{ref\_key}_{ij} = (s_i/i, r_j/j)$ and both words s_i and r_j appear in thinking process, then the $num\ of\ A^{ref\_key}\ hit\ in\ think$ should be incremented by one.

Overall Reward. Given the above rewards, the overall reward we design can be denoted as:

$$R_{all} = \begin{cases} 0, & \text{if } R_{format} = 0 \\ R_{comet} + \alpha * R_{aaw} & \\ \quad +\beta * R_{aao} & \\ \quad +\gamma * R_{taw}, & \text{if } R_{format} = 1 \end{cases} \quad (11)$$

where the hyperparameters α, β and γ control the trade-off between different reward components.

2.2 RL Algorithm

Our translation model is trained using the Group Relative Policy Optimization (GRPO) method (Shao et al., 2024), which optimizes policies through a hybrid reward function proposed in Section Design of Rewards. During training, for each input question q, we generate a set of candidate outputs $\{o_1, o_2, ..., o_G\}$ from the current policy model $\pi_{\theta_{old}}$. The advantage value A_i for each candidate is calculated by normalizing its reward r_i against the group's mean and standard deviation:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})} \quad (12)$$

GRPO then optimizes the policy parameters θ by maximizing the following objective:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \Bigg[\frac{1}{G}\sum_{i=1}^{G} \min\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \text{clip}\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\varepsilon, 1+\varepsilon \Big) A_i \Big) - \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) \Bigg] \tag{13}$$

3 Experiments and Results

In this section, we will introduce the relevant experimental setup, present the corresponding experimental results, and provide ablation studies.

3.1 Experimetal Setups

Backbone. We choose Qwen2.5-7B-Instruct (Yang et al., 2024) as the backbone model because it demonstrates strong multilingual performance among open-source models of comparable parameter size. This helps minimize potential negative impacts caused by insufficient capabilities of the base model. We also choose six models as our baselines, including GPT-4o (OpenAI, 2024a), Deepseek-V3 (DeepSeek-AI, 2025), Deepseek-R1 (DeepSeek-AI, 2025), Gemma-2-9B-it (Rivière et al., 2024), Llama-3.1-8B-Instruct (Meta, 2024) and TowerInstruct-7B-v0.2 (Alves et al., 2024).

Datasets. Following MT-R1-Zero (Feng et al., 2025), we used Chinese (ZH) to and from English (EN) parallel data from WMT 2017 to WMT 2020 as our training data. Additionally, we incorporate ZH-EN translation pairs from Flores-200 (Costa-jussà et al., 2022) and NTREX (Federmann et al., 2022), resulting in 16,124 training samples.

We selecte the ZH-EN test sets from WMT23 and WMT24 for general translation quality evaluation. For terminology-specific translation quality evaluation, we adopte the RTT test set (Zhang et al., 2023), a challenging English→German terminology test set containing 500 sentence pairs.

Evaluation Metrics. For general translation quality evaluation, we choose BLEU (Papineni et al., 2002; Post, 2018), COMETKiwi-23-XL (Rei et al., 2022), and XCOMET-XL (Guerreiro et al., 2024). BLEU is a lexical metric. COMETKiwi is a reference-free learning-based metric. XCOMET is a reference-based learning metric. These three metrics complement each other to some extent, enabling the evaluation of translation quality at both the lexical and semantic levels.

For terminology translation quality evaluation, in addition to the three metrics mentioned above, we also assess terminology accuracy (TA), indicating how many of the source terms have a corresponding target term in the translation.

Word Alignment. We select the open-source model SimAlign (Sabet et al., 2020) to extract word alignment information between the source text and the translation. SimAlign is an unsupervised word alignment tool with strong performance in word alignment tasks across multiple language pairs, including English and Chinese. SimAlign takes the tokenized sequences of the texts to be aligned as input. For Chinese, we use Jieba[2] for word segmentation, while for English, we employ NLTK[3].

Training Details. We conduct our training based on the verl [4] framework. For the hyperparameters in overall reward, we set α, β, and γ to 1, $\frac{1}{10}$, and $\frac{1}{10}$, respectively. In the GRPO algorithm, we set the number of rollouts to 16, the sampling temperature to 1.0, and use a constant learning rate of 1e-6. The maximum generation length is 4,096 tokens, with a training batch size of 128. All experiments are trained for three epochs.

3.2 Results and Analysis

This section presents the main experimental results, demonstrating that our proposed word-alignment reward is highly effective. We then provide a detailed analysis of the experimental outcomes and supplement the findings with relevant ablation studies.

Main Results. Table 1 presents the performance of models trained under different settings on the WMT test set, reflecting their general Chinese-English translation capabilities. Table 2 shows the results on the RTT test set, which demonstrate the models' terminology translation abilities. Here, *SFT* denotes the model obtained by fine-tuning Qwen2.5-7B-Instruct with our training data, while *RL-x* represents the model trained through reinforcement learning on Qwen2.5-7B-Instruct using our data, where *x* indicates different rewards employed during the reinforcement process. For example, RL-$R_{comet} + R_{BLEU}$ refers to the model reinforced using both COMET and BLEU as rewards with equal weights. The "NER + prompt" in Table 2

[2] https://github.com/fxsjy/jieba
[3] https://www.nltk.org/
[4] https://github.com/volcengine/verl

Models	ZH->EN				EN->ZH			
	BLEU	COMETKiwi	XCOMET	Avg.	BLEU	COMETKiwi	XCOMET	Avg.
GPT-4o	24.86	79.01	89.26	64.38	40.69	75.92	79.17	65.26
DeepSeek-V3-0324	23.87	78.89	89.12	63.96	36.25	76.59	80.92	64.59
Deepseek-R1-0210	22.93	78.27	89.11	63.44	36.61	75.79	80.14	64.18
Gemma-2-9b-it	22.00	75.40	87.85	61.75	39.48	71.19	76.17	62.28
TowerInstruct-7B-v0.2	23.36	72.45	85.41	60.41	32.89	66.89	71.83	57.20
Llama-3.1-8B-Instruct	18.47	58.94	83.40	53.60	32.44	66.41	72.54	57.13
Qwen2.5-7B-Instruct	22.18	73.08	86.05	60.44	37.36	71.65	75.46	61.49
SFT	22.15	73.98	86.27	60.80	33.42	68.58	75.19	59.06
RL-R_{comet}	22.32	77.51	88.59	62.80	36.12	75.79	79.42	63.78
RL-$R_{comet} + R_{BLEU}$	25.08	75.83	87.62	62.84	40.98	71.33	77.08	63.13
RL-$R_{comet} + R_{aaw}$	23.90	77.39	88.37	63.22	39.05	73.52	78.51	63.69
RL-$R_{comet} + R_{aaw} + R_{aao}$	23.97	77.21	88.27	63.15	38.53	74.94	78.65	64.04
RL-R_{all} (TAT-R1)	24.40	77.20	88.38	**63.33**	39.45	75.57	78.65	**64.56**

Table 1: Performance on WMT23 ZH to EN and WMT24 EN to ZH testset. ZH represents Chinese and EN represents English. *Avg.* represents the average of BLEU, COMETKiwi and XCOMET metrics.

Models	EN->DE				
	BLEU	COMETKiwi	XCOMET	TA	Avg.
Qwen2.5-7B-Instruct	25.87	67.05	88.65	53.29	58.72
NER + prompt	25.93	67.11	88.74	53.35	58.78
RL-R_{comet}	24.52	70.26	90.17	54.42	59.84
RL-R_{comet}+R_{BLEU}	27.37	66.49	88.77	54.91	59.39
RL-R_{comet}+R_{aaw}	26.21	71.33	90.56	55.57	60.92
RL-R_{comet}+R_{aaw}+R_{aao}	26.34	72.04	90.99	55.73	61.28
RL-R_{all} (TAT-R1)	27.10	**73.82**	**91.22**	**56.42**	**62.14**

Table 2: Performance on RTT testset. DE represents the German language. *Avg.* represents the average of BLEU, COMETKiwi, XCOMET and TA metrics.

represents the method which first extract the terminology using NER and then translate by prompt engineering. "Avg." in the table represents the average value of all metrics.

Regarding general translation performance (Table 1), our model TAT-R1 significantly improves across all metrics compared to the baseline Qwen2.5-7B-Instruct, and achieves SOTA performance among models of similar size, though it still lags behind closed-source large models like GPT-4o. For ZH→EN translation, the average metric increased from 60.44 to 63.33 (2.99% improvement), while for EN→ZH, it rises from 61.49 to 64.56 (3.07% improvement). Compared to using only COMET as a reward (RL-R_{comet}), incorporating three word alignment-related rewards further enhanced the model's overall performance metrics. Qwen2.5-7B-Instruct, the baseline we selected, is not a weak baseline. Its performance is only slightly lower than the 9B-parameter Gemma and significantly better than Llama-3.1-8B and TowerInstruct-7B-v0.2. Our model, TAT-R1, achieves SOTA performance among models of similar size, though it still lags behind closed-source large models like GPT-4o.

As shown in Table 2, on the terminology test set RTT, our model TAT-R1 with word alignment rewards demonstrates significant improvements over RL-R_{comet} (without word alignment rewards) across all evaluation metrics: BLEU score increased by 2.58%, COMETKiwi by 3.56%, XCOMET by 1.05%, and terminology accuracy (TA) by 2%.

Compared to models not reinforced with word alignment information, our TAT-R1 achieves slightly better performance in general translation tasks and significantly superior results in terminology translation, demonstrating the effectiveness of our proposed word alignment reward mechanism.

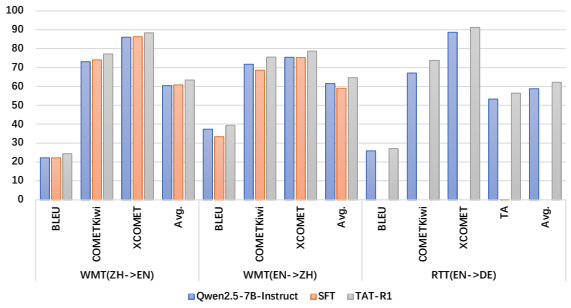

Figure 2: Compare the performance between SFT and RL.

Comparison between SFT and RL. To demonstrate the effectiveness of RL, we fine-tune the model using the same training data with SFT. As shown in Figure 2, although the fine-tuned model showed a slight improvement over the baseline in Chinese-to-English (Zh→En) translation on WMT, there was a noticeable decline in English-to-Chinese (En→Zh) metrics. We attribute this to noise in the current training data and that not all reference translations are of higher quality than

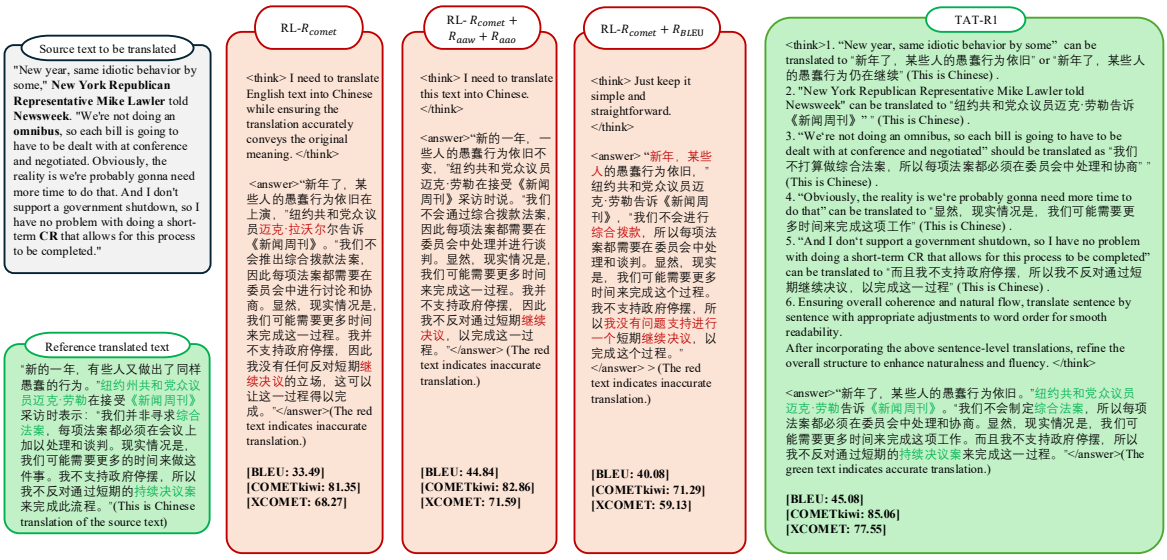

Figure 3: Qualitative examples illustrate the effect of different rewards on EN to ZH translation.

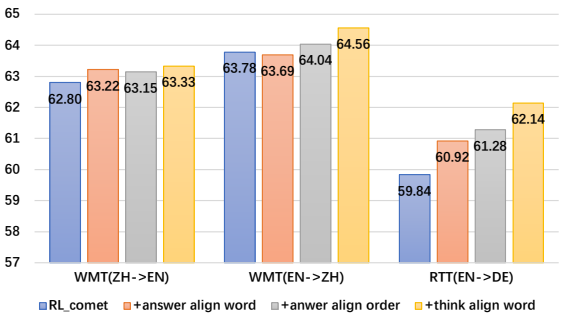

Figure 4: Compare the average performance between different word alignment rewards.

Datasets	Metrics	RL comet	RL comet+bleu	TAT-R1
WMT (ZH->EN)	BLEU	22.32	25.08	24.40
	COMETKiwi	77.51	75.83	77.20
	XCOMET	88.59	87.62	88.38
	Avg.	62.80	62.84	63.33
WMT (EN->ZH)	BLEU	36.12	40.98	39.45
	COMETKiwi	75.79	71.33	75.57
	XCOMET	79.42	77.08	78.65
	Avg.	63.78	63.13	64.56
RTT (EN->DE)	BLEU	24.52	27.37	27.10
	COMETKiwi	70.26	66.49	73.82
	XCOMET	90.17	88.77	91.22
	TA	54.42	54.91	56.42
	Avg.	59.84	59.39	62.14

Table 3: Compare between BLEU and word alignment rewards.

the model's original outputs, negatively impacting the translation performance after SFT. On the terminology test set RTT, the fine-tuned model almost entirely mistranslates English into Chinese for the English-to-German (En→De) task, resulting in all metrics dropping close to zero. In contrast, the RL-trained TAT-R1 model improved across all metrics, demonstrating strong performance on the out-of-distribution (OOD) En→De task. This phenomenon indicates that, in translation tasks, RL-trained models exhibit better stability and generalization capabilities compared to SFT.

The Effects of Different Word Alignment Rewards. Figure 4 demonstrates the average performance of the model when incrementally incorporating answer-align-word reward, answer-align-order reward, and think-align-word reward based on RL-R_{comet}. The results show that with the addition of each word alignment reward, the model's performance consistently improves, validating the effectiveness of our proposed word alignment rewards.

In Figure 3, we also present some output cases of the model after applying different rewards. We observe that when rewards are calculated only for the model's output within <answer> </answer> , the final "think" step often produces non-functional statements like "I need to translate the English text into Chinese and ensure the translation accurately conveys the original meaning."—failing to generate meaningful reasoning. This is evident in the RL-R_{comet}, RL-$R_{comet} + R_{aaw}$, and RL-$R_{comet} + R_{BLEU}$ examples in Figure 3. However, after introducing the think word alignment reward, the model begins to reason about the translation of key information in the "think" step, leading to a significant improvement in the final metrics, as

shown in the TAT-R1 example in Figure 3.

Comparison between BLEU and Word Alignment Rewards. As shown in Table 3, while the BLEU reward significantly improves the BLEU metric, it has a notably negative impact on semantic evaluation metrics such as COMET. We further analyze specific cases (e.g., RL-$R_{comet} + R_{BLEU}$ in Figure 3) and find that models trained with BLEU as the reward exhibit an apparent degradation in translation fluency. In contrast, the word alignment rewards focus solely on the correctness of keyword translations, demonstrating positive effects on lexical and semantic translation quality.

4 Related Work

4.1 Reason-based LLMs

In recent years, reason-based large language models, such as OpenAI's o1 (OpenAI, 2024b) and DeeeSeek-R1((DeepSeek-AI, 2025)), have demonstrated strong performance across various tasks, attracting significant attention from researchers. Recent studies primarily focus on solving complex reasoning tasks, such as mathematical problem-solving and code generation (Zeng et al., 2025; Hu et al., 2025; Luo et al., 2025; Song et al., 2025; Qin et al., 2024; Zhang et al., 2024). However, recent efforts have increasingly explored applying reason-based LLMs to general tasks. For instance, marco-o1 (Zhao et al., 2024) investigates the use of reasoning-enhanced models in open-ended text generation, where there are no clear-cut standards for evaluating correctness, unlike in mathematics or programming. Some surveys (Chen et al., 2025b; Li et al., 2025) provide systematic reviews of the advancements and trends in reason-based LLMs.

4.2 Reason-based LLMs for MT

Some researchers have attempted to explore the capabilities of reason-based LLMs in machine translation tasks. Marco-o1 (Zhao et al., 2024) and Liu et al. (2025) briefly demonstrate that reason-based LLMs can somewhat improve translation performance. DRT (Wang et al., 2024) enhances the model's effectiveness in literary translation by synthesizing translation data with reasoning processes and performing supervised fine-tuning (SFT). Chen et al. (2025a) provide a preliminary assessment of the performance of multiple reason-based LLMs in machine translation. Inspired by DeepSeek-R1 (DeepSeek-AI, 2025), some studies have tried to use reinforcement learning to stimulate the model's deep reasoning capabilities and improve translation quality. R1-T1 (He et al., 2025) synthesizes training data with reasoning processes for translation, first applying SFT and then conducting reinforcement training using COMET as the reward. Like DeepSeek-R1-Zero, MT-R1-Zero (Feng et al., 2025) directly performs reinforcement training on a pretrained model, employing BLEU and COMET as rewards. DeepTrans (Wang et al., 2025) directly uses DeepSeek-V3 scoring as the reward, enhancing the model's performance in literary translation through reinforcement learning.

4.3 Terminology Translation

In many fields, accurate translation of terminology is crucial. In recent years, numerous researchers have explored terminology translation using LLMs. Kim et al. (2024) detect terms, constructs a terminology database, and provides term information via retrieval-augmented generation (RAG) before model translation. Moslem et al. (2023) synthesizes bilingual data containing terms, fine-tunes the model, and applies post-processing to correct terminology after translation. For technical terms, Myung et al. (2024) propose a parenthetical terminology translation method. DragFT (Zheng et al., 2024) employs few-shot examples to enhance translation performance in specialized domains. Bogoychev and Chen (2023) improves term translation by constraining incorrect terminology during decoding. To better evaluate models' terminology translation capabilities, Zhang et al. (2023) introduce a new terminology test set and examines the effects of various data augmentation methods on term translation.

5 Conclusion

In this work, we introduce TAT-R1, the first terminology-aware translation model trained with RL and word alignment. Empowered by word alignment in machine translation, we design three types of new rule-based rewards. Combining the word alignment rewards with format reward and comet reward, we train our model with GRPO. Experimental results demonstrate the effectiveness of TAT-R1. TAT-R1 significantly improves terminology translation accuracy compared to the baseline while maintaining comparable performance on general translation tasks.

Limitations

While TAT-R1 has achieves significant improvements in terminology translation accuracy, certain limitaions remain. The reasoning process we observes is relative simple, and we have not observed the complex reasoning process such as self-correction and verification, which is appeared in mathematical tasks. This discrepancy may reflect the differences between machine translation task and mathematical task or indicate the need for specialized design in machine translation tasks. Another limitation is that we have not systematically explore multiple translation evaluation metrics as potential rewards, such as BLEURT (Sellam et al., 2020), MetricX (Juraska et al., 2024), and GEMBA (Kocmi and Federmann, 2023). A promising future research direction would be to investigate diverse reward signals for translation quality assessment, combined with word-alignment-based rewards, to further validate their effectiveness in terminology translation tasks.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 890–896. Association for Computational Linguistics.

Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025a. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *CoRR*, abs/2502.11544.

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. *CoRR*, abs/2503.09567.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2504.10160.

Nuno Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet : Transparent machine translation evaluation through fine-grained error detection. *Trans. Assoc. Comput. Linguistics*, 12:979–995.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, Hao Yang, Boxing Chen, and Osamu Yoshie. 2025. R1-T1: fully incentivizing translation capability in llms via reasoning learning. *CoRR*, abs/2502.19735.

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

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for llm-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 636–642. Association for Computational Linguistics.

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, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 193–203. European Association for Machine Translation.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025. From system 1 to system 2: A survey of reasoning large language models. *CoRR*, abs/2502.17419.

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

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

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 902–911. Association for Computational Linguistics.

Jiyoon Myung, Jihyeon Park, Jungki Son, Kyungro Lee, and Joohyung Han. 2024. Efficient technical term translation: A knowledge distillation approach for parenthetical terminology translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1410–1427, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o system card. *CoRR*, abs/2410.21276.

OpenAI. 2024b. Openai o1 system card. *CoRR*, abs/2412.16720.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. O1 replication journey: A strategic progress report - part 1. *CoRR*, abs/2410.18982.

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 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 80 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1627–1643. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *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. *Preprint*, arXiv:2503.17287.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2024. Drt-o1: Optimized deep reasoning translation via long chain-of-thought. *CoRR*, abs/2412.17498.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 2025. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. `https://hkust-nlp.notion.site/simplerl-reason`. Notion Blog.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6029–6042. Association for Computational Linguistics.

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

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. *CoRR*, abs/2411.14405.

Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *Preprint*, arXiv:2402.15061.