DeepTrans: Deep Reasoning Translation via Reinforcement Learning

Jiaan Wang, Fandong Meng*, Jie Zhou
Pattern Recognition Center, WeChat AI, Tencent Inc, China
jawang.nlp@gmail.com {fandongmeng,withtomzhou}@tencent.com

Abstract

Recently, deep reasoning LLMs (*e.g.*, OpenAI o1 and DeepSeek-R1) have shown promising performance in various downstream tasks. Free translation is an important and interesting task in the multilingual world, which requires going beyond word-for-word translation. However, the task is still under-explored in deep reasoning LLMs. In this paper, we introduce **DeepTrans**, a deep reasoning translation model that learns free translation via reinforcement learning (RL). Specifically, we carefully build a reward model with pre-defined scoring criteria on both the translation results and the thought processes. The reward model teaches DeepTrans how to think and free-translate the given sentences during RL. Besides, our RL training does not need any labeled translations, avoiding the human-intensive annotation or resource-intensive data synthesis. Experimental results show the effectiveness of DeepTrans. Using Qwen2.5-7B as the backbone, DeepTrans improves performance by 16.3% in literature translation, and outperforms strong deep reasoning LLMs. Moreover, we summarize the failures and interesting findings during our RL exploration. We hope this work could inspire other researchers in free translation.[1]

1 Introduction

Recently, deep reasoning LLMs (OpenAI, 2024b; Guo et al., 2025) have shown remarkable performance in tasks ranging from math (Chen et al., 2025b; Li et al., 2025), to coding (Zhang et al., 2024), question-taking (Guan et al., 2025), etc.

Some researchers bring the success of deep reasoning LLMs to neural machine translation (MT), with a particular focus on free translation rather than word-for-word translation (Zhao et al., 2024; Wang et al., 2025; Liu et al., 2025). Free translation is a translation approach that allows more flexibility to adapt the text to the target language, *taking into account* cultural nuances and making the text more natural and understandable for the target audience (Barbe, 1996; Chen et al., 2018). Given that, free translation is well-suited for deep reasoning LLMs to perform. Marco-o1 (Zhao et al., 2024) aims to extend deep reasoning LLMs into general domains where clear standards may be lacking. It uses illustrative examples to show the effectiveness of long chain-of-thought (CoT) in colloquial and slang MT. Wang et al. (2025) further systematically study the long CoT in literature MT. They find that literature sentences with metaphors or similes generally require cultural background to understand. Based on this insight, Wang et al. (2025) propose deep reasoning translation (DRT) models to translate the literature text from English to Chinese with step-by-step reasoning. The models are trained from synthesized long CoT translation data. Chen et al. (2025a) conduct empirical studies on how deep reasoning LLMs work on MT. They verify the strengths of deep reasoning LLMs in historical and cultural translation, but also point out their issues, *e.g.*, LLMs do not follow the instruction and fail to translate. More recently, R1-T1 (He et al., 2025) is presented, which is the first attempt to employ reinforcement learning (RL) in deep reasoning LLMs. Specifically, it uses COMET (Rei et al., 2020) score as the reward signal to optimize MT LLMs via modified REINFORCE++ (a RL training algorithm) (Hu, 2025). MT-R1 (Feng et al., 2025) uses a combination of BLEU and CometKiwi as the reward signal to optimize MT LLMs, and adopts GRPO (Shao et al., 2024) as the RL algorithm.

Meanwhile, RL has been verified to have a strong ability in deep reasoning LLMs, and LLMs can be equipped with powerful reasoning ability solely based on RL (Guo et al., 2025; Yu et al., 2025). However, RL is still under-explored in deep reasoning MT. In detail, Marco-o1 and DRT are only trained via supervised fine-tuning (SFT).

*Corresponding author.

[1] https://github.com/krystalan/DRT.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 47–63, 2026. https://doi.org/10.1162/TACL.a.65
Action Editor: Derek F. Wong. Submission batch: 4/2025; Revision batch: 8/2025; Published 1/2026.

Though R1-T1 and MT-R1 adopt RL training, using the COMET and CometKiwi as the rewards is not well calibrated. As pointed out by Liu et al. (2025), deep reasoning MT LLMs might achieve better translation but a lower COMET. Additionally, in literature domains, COMET and CometKiwi lose their effectiveness as evaluation metrics and show a poor correlation with human judgment (Karpinska and Iyyer, 2023; Wang et al., 2025).

In this paper, our research goal is to explore *how to improve the free translation ability of deep reasoning LLMs via reinforcement learning*. Following Wang et al. (2025), our explorations focus on the literature domain, where free translation is essential for addressing cultural differences. Subsequently, we focus on reward modeling, and study *how to employ an effective reward model during RL training*. There are three mainstream types of reward models in previous work: i) using MT metrics such as BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and CometKiwi (Rei et al., 2022); ii) training a reward model with preference data (Stiennon et al., 2020; Nakano et al., 2021), and using the reward model to infer a scalar reward during RL training; and iii) designing rule-based rewards (Guo et al., 2025). However, these rewards might lose their effectiveness or require high-quality annotated data in MT, hindering their usage. In detail, for i), the reference-based BLEU and COMET are unsuitable for literature MT since the high-quality translation data is difficult to collect. Though CometKiwi is a reference-free metric, its effectiveness is limited in the literature domain (Wang et al., 2025). For ii), training a reward model requires large-scale preference data, which is difficult to collect due to the expensive annotation costs. For iii), the rule-based rewards are suitable for tasks whose answers are easy to verify, *e.g.*, math and code problems (Swamy et al., 2025). In MT, we cannot design simple rules to judge the quality of translations.

In view of the strong ability of LLM-as-a-judge (Wang et al., 2023; Kocmi and Federmann, 2023; Li et al., 2024), we decide to use an advanced LLM, *i.e.*, DeepSeek-v3 (671B) (Liu et al., 2024), as the reward model. Specifically, we carefully design pre-defined scoring criteria on both the translation results and the thought processes. For the generation of deep reasoning MT LLMs, the reward model takes the pre-defined criteria into account to provide a discrete reward (*e.g.*, 3-point or 100-point scores). In this manner, we avoid collecting human-annotated preference data for training the reward model. In addition, the effectiveness of the reward model could also be ensured when we adopt the state-of-the-art LLM. We conduct extensive experiments on literature MT, using the above RL strategy to train DeepTrans-7B (with the backbone of Qwen2.5-7B). Experiments show the effectiveness of our method, DeepTrans-7B improves performance by 16.3% and outperforms strong deep reasoning baselines (*e.g.*, QwQ-32B-preview). Additionally, only training with source sentences, DeepTrans-7B outperforms DRT-7B (which is trained on synthesized long thought and translation data) in terms of both automatic metrics and GPT-4o evaluation. Moreover, we share the failures in our pilot RL experiments and summarize the critical component in reward modeling.

Our contributions are as follows:

- We propose DeepTrans, which aims to enhance the free translation ability of deep reasoning LLMs via RL. In detail, we use an LLM as the reward model, and carefully design scoring criteria on both translations and thought processes.
- Experimental results verify the effectiveness of our DeepTrans. Only training with the source sentences, DeepTrans outperforms strong deep reasoning baselines and LLMs that are fine-tuned with synthesized MT data.
- We summarize the failures and critical components during the RL training to provide a deeper understanding of deep reasoning MT LLMs.

2 Related Work

Deep Reasoning LLMs. In recent years, deep reasoning LLMs have pioneered growing research in long CoT reasoning. Different from the non-reasoning-oriented LLMs, deep reasoning LLMs involve a more detailed, iterative process of exploration and reflection within a given problem space (Chen et al., 2025b; Li, 2025). Many studies investigate the mathematical reasoning, programming tasks, and multidisciplinary knowledge reasoning capabilities of deep reasoning LLMs, and achieve promising performance (Yu et al., 2024; Sun et al., 2025; Yax et al.,

2024; Li et al., 2025; Guan et al., 2025; Jin et al., 2025). Some researchers investigate the MT capability of deep reasoning LLMs. Zhao et al. (2024) and Liu et al. (2025) briefly show that long CoT reasoning helps the model to reach more idiomatic translations. DRT (Wang et al., 2025) is further proposed in literary translation, and it is trained with synthesized SFT data. More recently, R1-T1 (He et al., 2025) and MT-R1 (Feng et al., 2025) employ RL to improve the translation ability of deep reasoning LLMs. However, R1-T1 and MT-R1 use the traditional MT metrics, *i.e.*, BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and COMETKiwi (Rei et al., 2022), as the RL reward. Different from them, we employ an advanced LLM with pre-defined scoring criteria to qualify the quality of both the translation and the thought process as the reward. In this way, the flaws of traditional metrics (Karpinska and Iyyer, 2023; Wang et al., 2025) can be avoided, and the strong ability of LLM-as-a-judge can be utilized to guide the RL training process effectively.

RL in Traditional MT. Machine translation via reinforcement learning has been explored before the deep reasoning LLM era. Early work tries to train MT models via optimizing BLEU scores (Ranzato et al., 2016; Shen et al., 2016; Bahdanau et al., 2017). Wu et al. (2016) design GLEU scores as the rewards to deal with the drawbacks of BLEU in single sentence evaluation. Wu et al. (2018) propose a method to involve large-scale monolingual data during RL training. Choshen et al. (2020) show the challenges of optimizing MT models via RL, *e.g.*, sparse reward signals and high-dimensional action space. Kiegeland and Kreutzer (2021) provide further analyses on these challenges. Kang et al. (2020) study document-level MT, and they propose a new method to both select context and translate sentences via RL.

3 DeepTrans

In this section, we introduce DeepTrans. As illustrated in Figure 1, there are three types of rewards: *format reward*, *thought reward* and *translation reward*. We first discuss the rewards designed in the RL framework (§ 3.1) and then provide the training details of DeepTrans (§ 3.2).

3.1 Reward Modeling

Given a source sentence, DeepTrans first thinks about how to translate the sentence, and then provides the translation result. We design the format of model generation as ‘‘`<think>` [thought] `</think>` [translation]’’, where ‘‘`<think>`’’ and ‘‘`</think>`’’ are two special tokens to indicate the boundary of thought content. ‘‘[thought]’’ and ‘‘[translation]’’ denote the content of thought and translation, respectively. Based on the model generation, we design the following rewards:

***Format Reward*:** We use a regular expression to judge whether the generation format is correct. In pilot experiments, we find that there might be some explanations in the translation results. To avoid it, we employ DeepSeek-v3 (Liu et al., 2024) to judge whether the translation results only contain translations. The judgment prompt is shown as follows:

> A translation question requires translating a given text from [src lang] into [trg lang].
>
> The given text is as follows:
> <text>
> {src}
> </text>
>
> Someone did this translation task and the translation result is as follows:
> <translation>
> {trans}
> </translation>
>
> Please judge whether the translation result belongs to the following situations:
> 1. It contains only the translation result.
> 2. It contains the translation result and the explanation.
> 3. It does not contain the translation result, but only the explanation.
>
> Please directly output your judgment result, such as: ‘‘Judgment result: 1’’, ‘‘Judgment result: 2’’ or ‘‘Judgment result: 3’’

where ‘‘{src}’’ and ‘‘{trans}’’ denote the source sentence and the translation result, respectively.

If both (a) the generation format is correct (determined by the regular expression), and (b) the translation result does not contain any explanations (determined by DeepSeek-v3), we regard the format as correct; otherwise, it is incorrect:

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

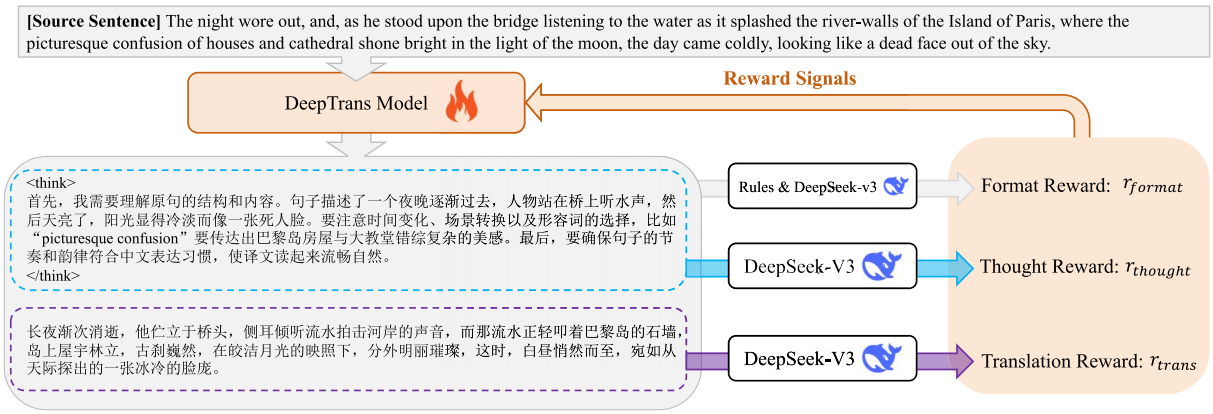


Figure 1: The overview of DeepTrans RL training.

Thought Reward: The thought process is important to guide the final translation. To reward meaningful thought processes, we adopt DeepSeek-v3 to provide the thought reward with a 3-point scale:

> A translation question requires translating a given text from [src lang] into [trg lang].
>
> The given text is as follows:
> <text>
> {src}
> </text>
>
> Someone did this translation question, and began to think how to translate:
> <think>
> {think}
> </think>
>
> Please judge whether there is a detailed analysis of the given text in this thinking process:
> 1. A lack of analysis: Only very shallow thinking was done, and no detailed analysis of the given text was carried out.
> 2. Slight analysis: The given text was analyzed in detail, and how to translate it was discussed in detail.
> 3. Detailed analysis: The given text was analyzed in detail, and various translation possibilities were discussed in detail, and trade-offs were made.
>
> Please directly output your judgment results, such as: ‘‘a lack of analysis’’, ‘‘slight analysis’’ or ‘‘detailed analysis’’

where ‘‘{think}’’ denotes the thought content. Subsequently, we define the thought reward as:

$$r_{\text{thought}} = \begin{cases} 2 & \text{if} v3^{\text{th}}(\text{src}, \text{think}) = \text{detailed analysis} \\ 1 & \text{if} v3^{\text{th}}(\text{src}, \text{think}) = \text{slight analysis} \\ 0 & \text{if} v3^{\text{th}}(\text{src}, \text{think}) = \text{a lack of analysis} \end{cases} \quad (2)$$

where $\text{v3}^{\text{th}}(\cdot, \cdot)$ denotes using DeepSeek-v3 as the thought reward scorer.

Translation Reward: We also employ DeepSeek-v3 to assess the translation quality. We borrow the prompt from Wang et al. (2025), which is designed to evaluate the quality of literature translation:

> SYSTEM PROMPT:
>
> Please evaluate the following Chinese translation of an English 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 and awkward phrasing that greatly hinder comprehension for a Chinese reader.
> - 30 points: Fair translation; the text conveys the basic meaning but lacks fluency and contains several awkward phrases or inaccuracies, making it challenging for a Chinese reader to fully grasp the intended message.
> - 50 points: Good translation; the text is mostly fluent and conveys the original meaning well, but may have minor awkwardness or slight inaccuracies that could confuse a Chinese reader.
> - 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 for a Chinese reader.
> - 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 for a Chinese reader.
>
> Please provide the reason first, followed by a score. Format your evaluation in the JSON structure below:
> {‘‘reason’’: ‘‘reason for the score’’, ‘‘score’’: int}

```
USER PROMPT:

<text>
{src}
</text>
<translation>
{trans}
</translation>
```

In this way, DeepSeek-v3 will generate its judgment for the translation result and provide a reward score on a 100-point scale:

$$r_{\text{trans}} = \text{v3}^{\text{tr}}(\text{src}, \text{trans}) \tag{3}$$

where $\text{v3}^{\text{tr}}(\cdot,\cdot)$ denotes using DeepSeek-v3 as the translation reward scorer.

***Overall Reward*:** Given the above three types of rewards, we finally design the overall reward:

$$r_{\text{all}} = \begin{cases} 0 & \text{if } r_{\text{format}} = 0 \\ r_{\text{trans}} + \alpha \times r_{\text{thought}} & \text{if } r_{\text{format}} \neq 0 \end{cases} \tag{4}$$

where α is a trade-off hyperparameter between r_{trans} and r_{thought}. Note that DeepSeek-v3 is used in r_{trans} / r_{thought}, and its effectiveness in MT evaluation is also verified by Larionov et al. (2025).

3.2 Training Details

Cold Start SFT. We use a non-reasoning LLM, *i.e.*, Qwen2.5-7B-Instruct (Yang et al., 2024), as the backbone of DeepTrans. To adapt DeepTrans to the deep reasoning MT, we first use DeepSeek-R1 (Guo et al., 2025) to generate seed translation samples in the general domain (*instead of the literature domain*). These samples follow our designed format, and are used to few-shot SFT the backbone model (named cold start SFT). The goal of cold start SFT is not to teach the model free translation, but the general translation with long thought.

RL Training. In view of the strong ability of GRPO (Shao et al., 2024), we adopt it in our RL training. For the policy model π, given it a source sentence s, GRPO first samples a number of generations $\{g_1, g_2, \ldots, g_n\}$ based on π, where each g_i involves a thought process and the translation result. Then, GRPO optimizes the policy model π' by maximizing the following objective:

$$\frac{1}{n}\sum_{1}^{n}(\min(\nabla_\pi A_i, \text{clip}(\nabla_\pi, 1-\epsilon, 1+\epsilon)A_i) - \beta\mathcal{D} \tag{5}$$

$$\nabla_\pi = \frac{\pi'(g_i|s)}{\pi(g_i|s)} \tag{6}$$

$$\mathcal{D} = \mathbb{D}_{kl}(\pi'||\pi_{ref}) \tag{7}$$

where ϵ and β are hyperparameters. π_{ref} is the reference model, and $\mathbb{D}_{kl}(\pi'||\pi_{ref})$ indicates the KL divergence between π' and π_{ref}. A_i denotes the advantage that is calculated as follows:

$$A_i = \frac{r^i_{\text{all}} - \text{mean}(\{r^1_{\text{all}}, r^2_{\text{all}}, \cdots, r^n_{\text{all}}\})}{\text{std}(\{r^1_{\text{all}}, r^2_{\text{all}}, \cdots, r^n_{\text{all}}\})}. \tag{8}$$

where r^i_{all} denotes the overall reward of g_i.

4 Experiments

4.1 Experimental Setups

Data. We use MetaphorTrans (Wang et al., 2025) in experiments, which is an English-Chinese literature MT data involving 19K training, 1K validation, and 2K test samples. Each sample involves a source English sentence, the corresponding Chinese translation, and the thought process during translation. The source sentences are selected from English literature books, and generally contain metaphors or similes. The thought processes and translation results are synthesized via Qwen2.5-72B-Instruct (Yang et al., 2024). We only use the source sentences during RL training.

In addition to the MetaphorTrans test set, we purchase electronic copies of two complete literature books, and use them to evaluate deep reasoning MT models: (1) *The Essential O. Henry Collection* (by O. Henry) and (2) *Orbital* (by Samantha Harvey). Both books are rich in literary nuance, making it challenging for even human translators to achieve a free translation.

Metrics. Since (1) the references in MetaphorTrans are synthesized via LLMs, and are not verified by human translators; and (2) the golden references of *The Essential O. Henry Collection* and *Orbital* are missing, we adopt reference-free metrics in our experiments. Specifically, we use *CometKiwi* (Rei et al., 2022) to evaluate the model translations, which judges whether a translation conveys the semantics of the source sentence. Moreover, following Wang et al. (2025), we use evaluators implemented using GPT-4o in two reference-free manners, which we refer to as *GRF* and *GEA*, respectively. The evaluation prompt of *GRF* borrows from Kocmi and Federmann

(2023).[2] For *GEA*, the prompt mainly borrows from Wang et al. (2025), and we employ two variants of *GEA*, *i.e.*, *GEA100* and *GEA5*. The evaluation prompt of *GEA100* is the same as the translation reward illustrated in §3.1, while that of *GEA5* simply narrows the scoring scope of *GEA100* from a 100-point to a 5-point scale. Among them, *GRF* evaluates translations from a general perspective while *GEA5* and *GEA100* evaluate translations from a literary perspective. The effectiveness of GRF in general translation and GEA in literary translation is demonstrated by Kocmi and Federmann (2023) and Wang et al. (2025), respectively. Since *GRF*, *GEA5* and *GEA100* need the costs of OpenAI's API, we randomly select 400 samples from each test set to conduct evaluation.

Backbone. Given the high computation costs in RL, we try to use LLMs (<10B parameters) as the backbone. Among all LLMs, Qwen2.5-7B (Yang et al., 2024) and Llama3-8B (Grattafiori et al., 2024) are state-of-the-art choices. Wang et al. (2025) show that Qwen2.5-7B outperforms Llama3-8B in literature translation. Thus, we use Qwen2.5-7B as the backbone of DeepTrans.

4.2 Implementation Details.

Cold Start SFT. We randomly select 4K English sentences from WMT24[3] in the general domain. Then, DeepSeek-R1 is employed to translate these sentences from English to Chinese in a deep reasoning manner. The synthesized 4K samples are used to SFT DeepTrans, named, cold-start SFT. Llama-Factory framework (Zheng et al., 2024) is used during the SFT stage. We conduct experiments on 8×NVIDIA H20 GPUs (96G) with 1e-5 learning rate and 8 (8×1) batch size. DeepSpeed ZeRO-3 optimization (Rasley et al., 2020) is also used during SFT. We set the number of SFT epochs to 2, and it costs about 1 GPU hour.

RL Training. We use GRPO RL algorithm implemented by verl.[4] 2×8 H20 GPUs are used, where 8 GPUs are used to deploy DeepSeek-v3 (awq quantization) as the reward model, and another 8 GPUs are used to optimize the policy model. We set the batch size to 64, the learning rate to 1e-6, the rollout number to 8 and the rollout temperature to 0.6, and the KL loss coefficient to 1e-3. The number of training epochs is set to 2. Since the scales of r_{trans} and r_{thought} are different, we set the trade-off hyperparameter α in Eq. 4 to 20. The RL training costs 2K GPU hours.

4.3 Baselines

(1) General Non-reasoning LLMs. We leverage Llama-3.1-8B-Instruct[5] (Grattafiori et al., 2024), Qwen2.5-7B-Instruct,[6] Qwen2.5-14B-Instruct[7] (Yang et al., 2024), and GPT-4o (OpenAI, 2024a) as baselines.

(2) General Reasoning LLMs. QwQ-32B-preview,[8] QwQ-32B[9] (Team, 2025), Marco-o1-7B[10] (Zhao et al., 2024), DeepSeek-Qwen-7B,[11] DeepSeek-Llama-8B,[12] DeepSeek-Qwen-14B,[13] DeepSeek-Qwen-32B,[14] DeepSeek-R1[15] (Guo et al., 2025), and o1-preview (OpenAI, 2024b) are used as baselines.

(3) MT Non-reasoning LLMs. Wang et al. (2025) fine-tune three LLMs with only paired sentences of the MetaphorTrans training data (without thought). This setting allows LLMs to learn the mapping from source literature sentences to the corresponding Chinese translations directly. The fine-tuned LLMs are denoted as Llama-3.1-8B-SFT, Qwen2.5-7B-SFT, and Qwen2.5-14B-SFT.

(4) MT Reasoning LLMs. Wang et al. (2025) introduce DRT-7B,[16] DRT-8B,[17] and DRT-14B[18] models, which are fine-tuned on the whole MetaphorTrans training data. Given sentences, these LLMs could first reason and then translate.

[2] Please refer to Figure 1 in Kocmi and Federmann (2023).

[3] https://www2.statmt.org/wmt24/index.html.

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

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

[6] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

[7] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct.

[8] https://huggingface.co/Qwen/QwQ-32B-Preview.

[9] https://huggingface.co/Qwen/QwQ-32B.

[10] https://huggingface.co/AIDC-AI/Marco-o1.

[11] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B.

[12] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B.

[13] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B.

[14] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B.

[15] https://huggingface.co/deepseek-ai/DeepSeek-R1.

[16] https://huggingface.co/Krystalan/DRT-7B.

[17] https://huggingface.co/Krystalan/DRT-8B.

[18] https://huggingface.co/Krystalan/DRT-14B.

	Para.	Deep Reason.	Training on MetaphorTrans	Ckpt.
General Non-reasoning Baselines				
Llama-3.1-8B-Instruct	8B	×	×	🤗
Qwen2.5-7B-Instruct	7B	×	×	🤗
Qwen2.5-14B-Instruct	14B	×	×	🤗
GPT-4o	–	×	×	–
General Reasoning Baselines				
Marco-o1-7B	7B	✓	×	🤗
QwQ-32B-preview	32B	✓	×	🤗
QwQ-32B	32B	✓	×	🤗
DeepSeek-Qwen-7B	7B	✓	×	🤗
DeepSeek-Llama-8B	8B	✓	×	🤗
DeepSeek-Qwen-14B	14B	✓	×	🤗
DeepSeek-Qwen-32B	32B	✓	×	🤗
DeepSeek-R1	671B	✓	×	🤗
o1-preview	–	✓	×	–
MT Non-reasoning Baselines				
Llama-3.1-8B-SFT	8B	×	✓	–
Qwen2.5-7B-SFT	7B	×	✓	–
Qwen2.5-14B-SFT	14B	×	✓	–
MT Reasoning Baselines				
DRT-7B	7B	✓	✓	🤗
DRT-8B	8B	✓	✓	🤗
DRT-14B	14B	✓	✓	🤗
Our				
DeepTrans-7B	7B	✓	✓	

Table 1: Comparisons between baselines. Para.: Parameter; Reason.: Reasoning. Ckpt.: Checkpoint.

To make a deeper understanding of these baselines, we summarize their key features in Table 1.

4.4 Main Results

As shown in Table 2, the experimental results verify the effectiveness of DeepTrans-7B. Specifically, compared with all baselines (<30B parameters), DeepTrans-7B achieves state-of-the-art performance in most metrics, especially GEA5 and GRF. DeepTrans-7B outperforms its backbone (Qwen2.5-7B-Instruct) by 16.3%, 13.8%, and 9.0% in terms of GEA5, GEA100, and GRF on MetaphorTrans. As for strong baselines that are fine-tuned on MetaphorTrans (denoted with ''♢''), they use the source sentences and the synthesized literature translations. In contrast, DeepTrans-7B only uses the source sentences and achieves promising results, and thus avoids the data quality issue and potential bias in synthesized data.

To figure out the effects of the cold start SFT and the RL training, we also compare DeepTrans-7B with its two variants: (1) *DeepTrans-7B (Cold Start)* is only trained via the cold start SFT and does not incorporate RL training. In contrast, (2) *DeepTrans-7B (Direct RL)* is trained solely using RL training without the cold start SFT. The experimental results in Table 2 show that both variants underperform the original DeepTrans-7B across all metrics, demonstrating the rationality and the effectiveness of the two-stage training process. The cold start SFT lets the backbone LLM (*i.e.*, Qwen2.5-7B-Instruct) quickly learn the deep reasoning translation task, and brings MT performance improvements to the model. Without the cold start SFT, DeepTrans-7B (Direct RL) only achieves sub-optimal performance. After the cold start SFT, the RL training stage further guides the model to acquire effective translation strategies. The substantial improvements of DeepTrans-7B compared to DeepTrans-7B (Cold Start) verify the superiority of the RL training.

We also compare DeepTrans-7B with baselines (>30B or commercial LLMs). As shown in Table 3, DeepTrans-7B generally outperforms QwQ-32B-preview and DeepSeek-Qwen-32B, and achieves competitive results with QwQ-32B, DeepSeek-R1, GPT-4o and o1-preview. This result demonstrates the superiority of DeepTrans-7B, with only 7B parameters, which shows strong performance in deep reasoning MT.

4.5 Human Evaluation

We conduct human evaluation to further evaluate the performance of DeepTrans-7B, DeepTrans-7B (Cold Start), DRT-14B, and Qwen2.5-14B-SFT. We randomly select 200 samples from each test set, 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 (Flu.), semantic accuracy (Sem.), and literary quality (Lit.). Following Kiritchenko and Mohammad (2017) and Wang et al. (2025), evaluators are tasked with identifying the best and worst translations for each aspect. The result scores are calculated based on the percentage of times each model is selected as best minus the times it is selected as worst. Consequently, the final scores range from -1 (indicating the worst performance) to 1 (indicating the best performance). As shown in Table 4, DeepTrans-7B significantly outperforms the others, especially in literary quality. This result demonstrates the superiority and effectiveness of DeepTrans-7B. The Fleiss' Kappa scores (Fleiss, 1971) of Flu., Sem., and Lit. are

Model	MetaphorTrans				*O. Henry*				*Orbital*			
	GRF	GEA5	GEA100	CometKiwi	GRF	GEA5	GEA100	CometKiwi	GRF	GEA5	GEA100	CometKiwi
[1pt] Llama-3.1-8B-Instruct	79.25	3.31	59.58	70.14	79.73	3.43	57.17	74.35	79.92	3.54	59.90	75.09
Qwen2.5-7B-Instruct	81.53	3.62	66.21	70.36	85.26	3.83	66.50	76.18	83.38	4.00	70.10	76.46
Qwen2.5-14B-Instruct	84.74	3.87	70.86	72.01	86.83	3.98	70.53	77.04	84.39	4.09	71.65	76.55
Marco-o1-7B	82.41	3.57	64.24	71.62	83.12	3.71	63.11	76.00	81.84	3.89	67.64	75.38
DeepSeek-Qwen-7B	65.16	2.67	43.66	63.49	68.97	2.86	45.64	70.67	71.28	3.16	51.91	72.43
DeepSeek-Llama-8B	76.31	3.24	56.89	67.13	78.17	3.39	56.14	73.39	78.91	3.64	59.75	74.47
DeepSeek-Qwen-14B	83.92	3.81	70.64	71.01	83.27	3.82	64.79	75.22	82.30	4.01	69.10	76.28
Llama-3.1-8B-SFT$^\diamondsuit$	84.10	3.88	69.33	70.25	85.04	3.87	66.60	76.14	80.37	3.87	64.38	75.11
Qwen2.5-7B-SFT$^\diamondsuit$	85.06	3.93	72.29	71.03	86.84	4.05	71.05	77.29	85.46	4.12	70.55	76.32
Qwen2.5-14B-SFT$^\diamondsuit$	85.66	4.02	74.53	72.08	87.27	4.05	73.06	**77.54**	85.55	4.14	75.84	**77.40**
DRT-7B$^\diamondsuit$	85.57	4.05	75.05	71.78	86.36	3.96	69.51	76.12	81.69	3.84	65.56	69.95
DRT-8B$^\diamondsuit$	84.49	3.91	69.65	70.85	83.61	3.75	64.76	73.89	79.14	3.65	61.36	66.36
DRT-14B$^\diamondsuit$	87.19	4.13	**77.41**	**72.11**	87.38	4.00	72.59	76.70	82.19	3.98	69.36	70.99
DeepTrans-7B (Cold Start)	85.06	3.94	66.72	71.49	85.90	4.03	69.01	76.91	85.22	3.97	73.61	76.54
DeepTrans-7B (Direct RL)	87.13	4.11	71.43	70.68	86.37	4.01	72.93	75.97	85.92	4.09	74.14	76.00
DeepTrans-7B$^\diamondsuit$ (Our)	**88.84**†	**4.21**‡	75.38	71.82	**87.95**†	**4.22**†	**76.92**†	77.04	**87.95**†	**4.22**†	**76.92**†	76.65

Table 2: Comparison results of DeepTrans and open-source baselines (<30B parameters). The **bold** and the underline denote the best and second-best scores, respectively. ‘‘†’’ and‘‘‡’’ denote statistically significant better than the DRT-14B with t-test p < 0.01 and 0.05, respectively. ‘‘$\diamondsuit$’’ denotes models are trained on MetaphorTrans.

Model	MetaphorTrans				*O. Henry*				*Orbital*			
	GRF	GEA5	GEA100	CometKiwi	GRF	GEA5	GEA100	CometKiwi	GRF	GEA5	GEA100	CometKiwi
GPT-4o	85.57	3.86	71.88	73.01	88.30	4.00	71.06	76.74	85.91	4.17	73.54	77.67
o1-preview	87.11	4.06	**78.01**	**73.70**	89.73	4.14	76.17	**78.41**	86.85	4.26	76.80	**78.86**
QwQ-32B-preview	86.31	4.00	75.50	71.48	87.61	4.03	70.79	76.86	84.79	4.04	71.03	76.17
QwQ-32B	88.06	4.09	74.38	72.88	88.02	4.21	76.36	77.71	87.83	4.15	76.55	77.55
DeepSeek-Qwen-32B	84.78	3.87	71.88	71.93	87.03	4.03	70.81	76.75	85.36	4.16	73.62	77.80
DeepSeek-R1	84.29	4.02	73.78	68.33	**89.79**	4.17	**77.03**	77.01	87.37	**4.27**	**80.06**	76.17
DeepTrans-7B (Our)	**88.84**	**4.21**	75.38	71.82	87.95	**4.22**	76.92	77.04	**87.95**	4.22	76.92	76.65

Table 3: Comparison results of DeepTrans and baselines (>30B parameters or commercial LLMs).

Model	Flu.	Sem.	Lit.
MetaphorTrans			
Qwen2.5-14B-SFT	0.044	−0.009	−0.105
DRT-14B	0.079	0.079	0.109
DeepTrans-7B (Cold Start)	−0.258	−0.291	−0.269
DeepTrans-7B	**0.135**	**0.221**	**0.265**
O. Henry			
Qwen2.5-14B-SFT	−0.020	−0.019	−0.170
DRT-14B	0.088	0.072	0.129
DeepTrans-7B (Cold Start)	−0.245	−0.294	−0.241
DeepTrans-7B	**0.177**	**0.241**	**0.282**
Orbital			
Qwen2.5-14B-SFT	−0.010	0.001	−0.120
DRT-14B	0.091	0.082	0.109
DeepTrans-7B (Cold Start)	−0.238	−0.274	−0.241
DeepTrans-7B	**0.157**	**0.191**	**0.252**

Table 4: Human evaluation results in terms of fluency, semantic accuracy, and literary quality.

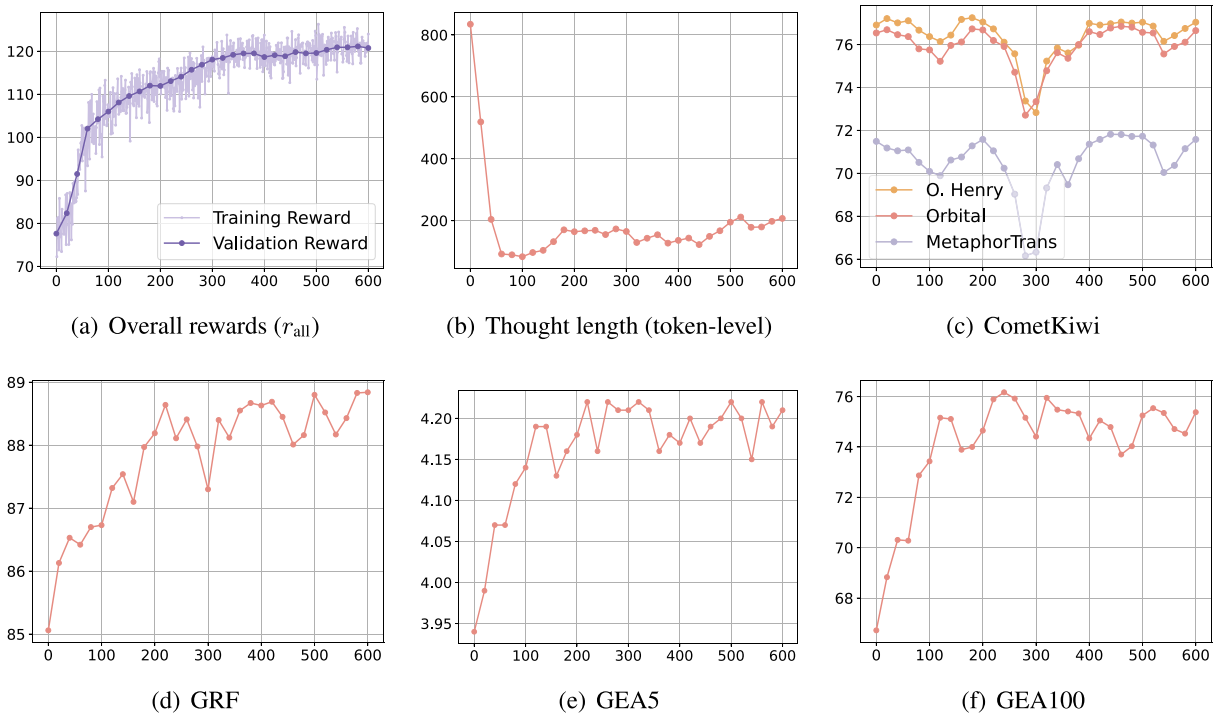
0.68, 0.70, and 0.74, respectively, indicating a good inter-agreement among evaluators.

4.6 Intermediate-Stage Analyses

To provide a deeper analysis of DeepTrans, we discuss the performance changes during RL training. Figure 2 shows the corresponding details, which we analyze from the following aspects:

In terms of *overall rewards* (c.f. Figure 2(a)), the training rewards and validation rewards generally increase along with the training process. The full score of the overall reward is 140 (according to Eq. 4), and DeepTrans-7B finally reaches about 120, which means r_{trans} reaches at least 80. According to the definition of r_{trans}, 80 points indicate the translations are between ‘‘very good translation’’ and ‘‘excellent translation’’, showing the superiority of DeepTrans-7B. Similarly, as shown in Figure 2(d), (e), and (f), model performance in GRF, GEA5 and GEA100 also generally increases along with the training process. After RL training, DeepTrans-7B demonstrates a substantial performance enhancement compared to the initial model, which is only trained via the cold start SFT. This also verifies the effectiveness of RL training.

In terms of *thought length* (c.f. Figure 2(b)), the average length first significantly decreases from 800+ to 100 tokens, and then slowly increases to 200 tokens. This finding indicates that DeepTrans-7B first generates plenty of thought content owing to the cold start SFT, and thus can obtain high thought rewards. During RL training, the model first preliminarily focuses on the translation results instead of the thought content, and it starts to simplify the thought content, since no additional thought rewards could be given. However, when the thought length is below a threshold, the



(a) Overall rewards (r_{all}) (b) Thought length (token-level) (c) CometKiwi

(d) GRF (e) GEA5 (f) GEA100

Figure 2: Performance changes of DeepTrans-7B during RL training. Among them, (c) is conducted on all three test sets, while others are conducted on MetaphorTrans. The horizontal axis denotes the number of training steps, and there are 600 steps (2 epochs) in total. The vertical axis denotes the value of the corresponding metrics.

Source Sentence	DeepTrans (Cold Start)	DeepTrans (150 Steps)	DeepTrans (300 Steps)	DeepTrans (450 Steps)	DeepTrans (600 Steps)
When it was light enough Johnsy, the merciless, commanded that the shade be raised.	当光线足够明亮时，无情的约翰西斯命令把窗帘拉上。 **[CometKiwi: 81.72]**	当光线足够明亮时，那个无情的约翰西便下令拉起窗帘。 **[CometKiwi: 81.56]**	待天色稍霁，约翰西这颗冷酷的心，便命将帘幕徐徐卷起。 **[CometKiwi: 57.61]**	等到光线足够了，这个冷酷无情的约翰西，便下令将那阴影缓缓卷起。 **[CometKiwi: 72.48]**	约翰西，这个冷酷无情的人，等到天色足够明亮，便命令将帘幕缓缓拉起。 **[CometKiwi: 77.28]**
When one loves one's Art no service seems too hard.	当一个人热爱自己的艺术时，任何服务都显得不值一提。 **[CometKiwi: 70.98]**	一旦钟爱艺术，便觉任何奉献皆非难事。 **[CometKiwi: 81.86]**	一旦钟爱自家的艺术，任是何般劳顿艰辛，也觉微不足道，甘之如饴。 **[CometKiwi: 65.53]**	只要钟爱自己的艺术，再大的付出也觉甘愿。 **[CometKiwi: 67.99]**	只要钟爱自己的艺术，再大的付出也绝不嫌苦。 **[CometKiwi: 76.16]**
For two or three days she went out canvassing for pupils. One evening she came home elated.	连续两三天她出去招生，一个傍晚她兴高采烈地回家了。 **[CometKiwi: 83.55]**	连续两三天，她外出物色学生。某晚归来时，她显得欣喜若狂。 **[CometKiwi: 84.96]**	接连三四日，她外出寻觅门徒踪迹，直至某晚归来，已得诸多佳音满怀欣喜。 **[CometKiwi: 64.59]**	连续一两天，她便出门去物色学生。一天傍晚，她兴冲冲地回到了家中。 **[CometKiwi: 82.87]**	连续两三天，她外出寻觅学生上门求教。一天傍晚，她满面春风地回到了家中。 **[CometKiwi: 81.21]**

Table 5: Case Studies on *O. Henry*. Purple indicates the translations are in strong Chinese classical literary style.

model cannot provide a meaningful thought content with short words, and it will be punished by r_{thought}. Then, the model will stop simplifying the thought content, and learn to generate better translations and keep high-quality thought processes simultaneously. Later, along with RL training, the model will find deeper thoughts that will lead to better translations, and learn to improve its thinking.

In terms of *CometKiwi* (c.f. Figure 2(c)), we find an interesting phenomenon: The performance drops a lot at first, and finally returns to the initial performance. To understand the reason behind this, we provide several cases generated by the intermediate-stage DeepTrans-series models. As shown in Table 5, we find that the model trained with 300 steps (*i.e.*, 1 epoch) has a special characteristic: **its translations exhibit a distinctly classical Chinese literary style** (*e.g.*, ancient poetry). We mark several classical-style terms in purple, and these terms are also not commonly used in modern Chinese. The classical style might

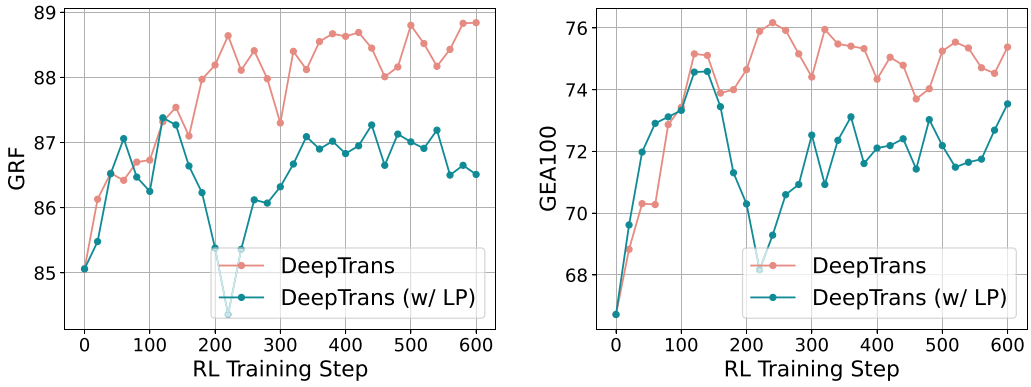


Figure 3: The comparisons between DeepTrans and DeepTrans (w/ LP) on MetaphorTrans.

not be well captured by the CometKiwi. However, it still can be recognized by the GPT-4o evaluators. As shown in Figure 2, DeepTrans (300 steps) still outperforms DeepTrans (cold start) in terms of GRF, GEA5, and GEA100. For this stylization phenomenon, we also want to discuss the following questions: *(1) Is stylization better or not?* Though classical-style translations will provide some interesting results, we think this might lose the generalizability of the MT models. The classical-style translations are not suitable for all genres in literature. Our goal is to make the model provide translations that should be more easily absorbed by native people in target languages. The classical-style translations do not follow this goal, and it should be considered a bias or preference. Thus, we recommend that future work should also evaluate the stylization phenomenon during RL training. *(2) Why does DeepTrans first drop into the stylization phenomenon and finally get out of it?* First, the model recognizes that the classical-style translations will receive a high translation reward, and thus, the model starts to provide classical-style translations. This also indicates a bias/preference of the reward model. However, such a stylization strategy will trap the model into a local optimal solution. Fortunately, the model discovers other better solutions during exploration in RL, and finally discards the stylization strategy. Another important question is naturally raised: *which factors lead the model to get out of the local optimal solution?* We will discuss it in §5.3.

5 Failure Experience

We discuss our failures in reward design, which could also be regarded as the ablation study of r_{all}.

5.1 The Length Penalty of Thought Process

In the pilot RL experiments, we design a length penalty for the thought content:

$$r_{\text{length}} = \max(-\frac{\text{len(think)} - \beta}{\eta}, 0) \tag{9}$$

where β and η are hyperparameters. len(think) indicates the token-level length of the generated thought content. In this manner, the length penalty will punish redundant thinking processes, and we add it to r_{thought}:

$$r_{\text{thought (w/ penalty)}} = r_{\text{thought}} + r_{\text{length}} \tag{10}$$

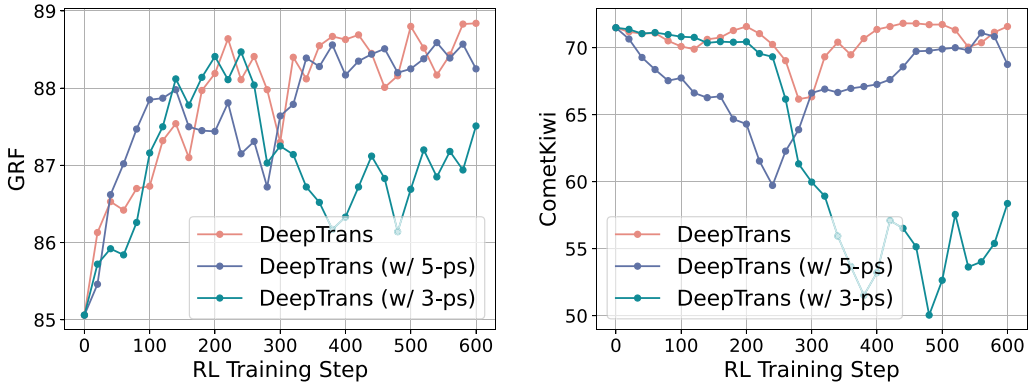
We set β to 400 and η to 400 in the pilot experiments. The DeepTrans-7B model trained with the length penalty is denoted as DeepTrans (w/ LP). As shown in Figure 3, we find that the length penalty does not bring improvement during training. In contrast, it reduces the rate of improvement in model performance. Therefore, we do not adopt the length penalty in the final reward modeling.

5.2 The Effect of Thought Reward

We also attempt to remove the thought reward, and the variant overall reward is defined as:

$$r_{\text{all (w/o th)}} = \begin{cases} 0 & \text{if } r_{\text{format}} = 0 \\ r_{\text{trans}} & \text{if } r_{\text{format}} \neq 0 \end{cases} \tag{11}$$

However, we find that DeepTrans trained via $r_{\text{all (w/o th)}}$ quickly discards the thought process, and tends to directly output the final translation until the end of the training. To make the model provide a valuable thought process, we further consider a short length penalty on the thought



(a) In terms of GRF (b) In terms of CometKiwi

Figure 4: Comparisons of models trained with translation rewards across different scoring scales. The results are evaluated on the MetaphorTrans test set.

process: If the thought length is less than k tokens, we give a huge penalty to the overall reward:

$$r_{\text{all(w/o th \& w/ sp)}} = \begin{cases} 0 & \text{if } r_{\text{format}} = 0 \\ 0 & \text{len(think)} < k \\ r_{\text{trans}} & \text{else} \end{cases} \quad (12)$$

We set k to 10, and find that the trained model tends to only provide several empty words in the thought content. For example, ‘‘Well, first I have to understand the meaning of the sentence. Then, I will express it in fluent Chinese’’. This kind of thought content loses relevance and specificity towards the source sentences.

Besides, the model trained with $r_{\text{all (w/o th)}}$ or $r_{\text{all (w/o th \& w/ sp)}}$ significantly underperforms the original DeepTrans-7B in terms of all metrics. These findings demonstrate the importance of the thought process in model application, and we should consider how to guide models to generate valuable thought during RL training.

5.3 The Effect of Translation Reward

We define the translation reward using a 100-point scale in §3.1, and we further study the effect of the scoring criteria on translation rewards. Specifically, we use two variants: $r_{\text{trans (3 point)}}$ and $r_{\text{trans (5 point)}}$ simply narrow the scoring scope of r_{trans} from a 100-point scale to a 3-point scale and a 5-point scale, respectively. During training with these two variants, the α in Eq. 4 is also changed to ensure the same ratio between thought rewards and translation rewards, *i.e.*, we set α to 0.6 and 1.0 when adopting $r_{\text{trans (3 point)}}$ and $r_{\text{trans (5 point)}}$, respectively. We denote the models trained with the variants as ‘‘DeepTrans (w/ 3-ps)’’ and ‘‘DeepTrans (w/ 5-ps)’’, where ‘‘ps’’ is the abbreviation of ‘‘point scale’’. As shown in Figure 4(a), both variants show the sub-optimal results in terms of GRF, verifying the rationality of our original design.

As for CometKiwi in Figure 4(b), we find that DeepTrans (w/ 5-ps) shows similar trends with original DeepTrans, *i.e.*, first drops and finally recovers. However, DeepTrans (w/ 3-ps) continues to decline and fluctuate during training. In the end, it achieves a low CometKiwi score, and perhaps it needs more training steps to recover the performance. In view of the stylization phenomenon we discussed in § 4.6, we conjecture the trends shown in Figure 4(b) are attributed to *stylization*.

To figure it out, we provide case studies on the variant models in Table 6. When the models are trained with 300 steps (1 epoch), both variants show the stylization phenomenon. Further, when trained with 600 steps (2 epochs), we find that DeepTrans (w/ 5-ps) discards the classical-style translations. In contrast, DeepTrans (w/ 3-ps) enhances the classical-style translations. As we also described in §4.6, the basic reason of the stylization phenomenon is a preference of the reward model, which encourages the MT model to provide stylized translations as a local optimal solution. When the reward model provides a wider scoring scope, it gives a larger latent space for the policy model to explore. Thus, the policy model is more likely to get out of the local optimal solution. If using a narrow reward scope, some minor changes in the policy model will not be reflected by the reward signal immediately, making it hard

Source Sentence	DeepTrans (Cold Start)	DeepTrans Variants (300 Steps)		DeepTrans Variants (600 Steps)	
		w/ 3-point scale	w/ 5-point scale	w/ 3-point scale	w/ 5-point scale
When it was light enough Johnsy, the merciless, commanded that the shade be raised.	当光线足够明亮时，无情的约翰西斯命令把窗帘拉上。 **[CometKiwi: 81.72]**	当光线已然充足，约翰西，那无情之人，竟令帷幕徐徐卷起。 **[CometKiwi: 74.46]**	待天色渐明，那冷漠无情的约翰西，便命人将帘幕徐徐卷起。 **[CometKiwi: 75.08]**	天色稍显明亮，约翰西，那颗铁石心肠，便下令将阴影卷起。 **[CometKiwi: 66.96]**	待天色足够明亮，约翰西这颗铁石心肠，已然下令将帘幕卷起。 **[CometKiwi: 71.40]**
When one loves one's Art no service seems too hard.	当一个人热爱自己的艺术时，任何服务都显得不值一提。 **[CometKiwi: 70.98]**	若钟情于艺道，任劳任怨亦无妨。 **[CometKiwi: 69.38]**	一旦钟爱所钟之艺，便觉任劳任怨非艰。 **[CometKiwi: 69.38]**	艺海倾心无难事，任凭千般苦亦甘。 **[CometKiwi: 54.76]**	只要钟爱所钟情的艺术，任何付出都觉轻而易举。 **[CometKiwi: 76.59]**
For two or three days she went out canvassing for pupils. One evening she came home elated.	连续两三天她出去招生，一个傍晚她兴高采烈地回家了。 **[CometKiwi: 83.55]**	连续几日外出寻觅门徒，一夕归家心欢悦。 **[CometKiwi: 80.25]**	她连着几日四处奔走寻觅门徒，直至某晚，她满心欢喜地踏入了家园。 **[CometKiwi: 74.54]**	连续几日奔走寻门徒，直至傍晚方归家。满心欢喜步履轻，笑颜如花映晚霞。 **[CometKiwi: 55.10]**	她接连走了两三天，到处奔走寻觅学生。一天傍晚，她满面春风地回到了家中。 **[CometKiwi: 83.87]**

Table 6: The results of case studies on *O. Henry*.

to transfer the translation strategies. In conclusion, *the scoring scope of the reward models has a great effect on the policy model, and it is necessary to employ a wider reward scope for the model to explore a good translation strategy.*

6 Conclusion and Future Work

In this work, we propose DeepTrans, which aims to enhance LLMs' deep reasoning MT ability via RL. We use DeepSeek-v3 as the reward model to provide discrete reward values. The experimental results in the literature MT show the effectiveness of DeepTrans. It outperforms QwQ-32B-preview, and achieves competitive results with state-of-the-art deep reasoning LLMs. In addition, we give deeper analyses on model results during RL training, and find a stylization phenomenon that needs to be carefully considered. Moreover, we summarize the failure experiences and critical components in the RL framework to provide a deeper understanding of deep reasoning MT LLMs.

In the future, the following directions may be worth exploring to promote the deep reasoning translation: (1) Extending deep reasoning translation models into multi-lingual MT. Consequently, the model could think and translate among different languages; (2) Qualifying and controlling the translation style during RL training to alleviate the stylization phenomenon (Section 4.6); (3) Exploring translation rewards with enhanced accuracy, robustness, or efficiency. In this work, we prompt vanilla DeepSeek-v3 to provide the translation reward, which is computationally intensive; (4) Balancing long- and short-thought to improve models' inference efficiency, and avoid the overthinking issue (Sui et al., 2025). For simple translation queries, a long thought process may not be necessary.

7 Limitations

While we show the effectiveness of reinforcement learning in deep reasoning translation, there are some limitations worth noting: (1) We evaluate the deep reasoning translation models in a single translation direction, *i.e.*, English-to-Chinese, and future work could extend our method to other directions; (2) The reward model employed in our experiments is DeepSeek-v3, which requires significant computational resources to infer; (3) Due to resource limitations, we conduct experiments on LLMs with 7B parameters, and future work could extend our method to LLMs with more parameters.

Acknowledgments

We thank the action editor and anonymous reviewers for their suggestions. These suggestions are very helpful for improving our work.

References

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017.

An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations*.

Katharina Barbe. 1996. The dichotomy free and literal translation. *Meta*, 41(3):328–337. https://doi.org/10.7202/001968ar

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. *ArXiv preprint*, abs/2502.11544v1.

Qi Chen, Oi Yee Kwong, and Jingbo Zhu. 2018. Detecting free translation in parallel corpora from attention scores. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Hong Kong. Association for Computational Linguistics.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. 2025b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *ArXiv preprint*, abs/2503.09567v3.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. 2020. On the weaknesses of reinforcement learning for neural machine translation. In *International Conference on Learning Representations*.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. MT-R1-Zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *ArXiv preprint*, abs/2504.10160v1. https://doi.org/10.18653/v1/2025.findings-emnlp.1015

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378. https://doi.org/10.1037/h0031619

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, et al. 2024. The Llama 3 herd of models. *ArXiv preprint*, abs/2407.21783v3.

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*, abs/2502.01142v1.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu,

Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv preprint*, abs/2501.12948v1. https://doi.org/10.1038/s41586-025-09422-z, PubMed: 40962978

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*, abs/2502.19735v2.

Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *ArXiv preprint*, abs/2501.03262v1.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training llms to reason and leverage search engines with reinforcement learning. *ArXiv preprint*, abs/2503.09516v3.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.175

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.41

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

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2074

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

Daniil Larionov, Sotaro Takeshita, Ran Zhang, Yanran Chen, Christoph Leiter, Zhipin Wang, Christian Greisinger, and Steffen Eger. 2025. DeepSeek vs. o3-mini: How well can reasoning llms evaluate mt and summarization? *ArXiv preprint*, abs/2504.08120v1.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and others. 2024. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *ArXiv preprint*, abs/2411.16594v6.

Xinzhe Li. 2025. A survey on LLM test-time compute via search: Tasks, LLM profiling, search algorithms, and relevant frameworks. *Transactions on Machine Learning Research*.

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. *ArXiv preprint*, abs/2502.17419v2.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, et al. 2024. DeepSeek-V3 technical report. *ArXiv preprint*, abs/2412.19437v2.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. New trends for modern machine translation with large reasoning models. *ArXiv preprint*, abs/2503.10351v1.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *ArXiv preprint*, abs/2112.09332v3.

OpenAI. 2024a. GPT-4o system card. *ArXiv preprint*, abs/2410.21276v1.

OpenAI. 2024b. Learning to reason with large language models. https://openai.com/index/learning-to-reason-with-llms/

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*.

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. https://doi.org/10.1145/3394486.3406703

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. https://doi.org/10.18653/v1/2020.emnlp-main.213

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa 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)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). 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. *ArXiv preprint*, abs/2402.03300v3.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Yuan Wu, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2025. A survey of reasoning with foundation models: Concepts, methodologies, and outlook.*ACM Computing Surveys*, 57(11):1–43.

Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J. Andrew Bagnell. 2025. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *ArXiv preprint*, abs/2503.01067v1.

Qwen Team. 2025. QwQ-32B: Embracing the power of reinforcement learning.

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. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.newsum-1.1

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025. DRT: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.351

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*, abs/1609.08144v2.

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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan,

Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al. 2024. Qwen2.5 technical report. *ArXiv preprint*, abs/2412.15115v2.

Nicolas Yax, Hernán Anlló, and Stefano Palminteri. 2024. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51. https://doi.org/10.1038/s44271-024-00091-8, PubMed: 39242743

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39. https://doi.org/10.1145/3664194

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: An open-source llm reinforcement learning system at scale. *ArXiv preprint*, abs/2503.14476v1.

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*, abs/2412.00154v2.

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*, abs/2411.14405v2.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-demos.38