DRT: Deep Reasoning Translation via Long Chain-of-Thought

Anonymous ACL submission

Abstract

Recently, O1-like models have emerged as representative examples, illustrating the effectiveness of long chain-of-thought (CoT) in reasoning tasks such as math and coding tasks. In this paper, we introduce DRT, an attempt to bring the success of long CoT to neural machine translation (MT). Specifically, in view of the literature books that might involve similes and metaphors, translating these texts to a target language is very difficult in practice due to cultural differences. In such cases, literal translation often fails to convey the intended meaning effectively. Even for professional human translators, considerable thought must be given to preserving semantics throughout the translation process. To simulate LLMs' long thought ability in MT, we first mine sentences containing similes or metaphors from existing literature books, and then develop a multi-agent framework to translate these sentences via long thought. In the multi-agent framework, a translator is used to iteratively translate the source sentence under the suggestions provided by an advisor. To ensure the effectiveness of the long thoughts, an evaluator is also employed to quantify the translation quality in each round. In this way, we collect tens of thousands of long-thought MT data, which is used to train our DRT. Using Qwen2.5 and LLama-3.1 as the backbones, DRT models can learn the thought process during machine translation, and outperform vanilla LLMs as well as LLMs which are simply fine-tuning on the paired sentences without long thought, showing its effectiveness.[1]

1 Introduction

Recently, the emergence of the O1-like LLMs shows great performance in reasoning tasks, *e.g.*, math and coding tasks (OpenAI, 2024b; Qin et al., 2024; Huang et al., 2024; Zhang et al., 2024; Zhao et al., 2024). With the help of long thought, LLMs tend to explore, reflect and self-improve the reasoning processes to achieve more accurate answers.

In this paper, we explore technical routes to bring the success of long thought to MT. To this end, we introduce DRT, a product of our exploration, and we hope it could facilitate the research community. There are two key points in achieving this goal:

i) **A suitable translation scenario to employ long thought in MT**: Not all scenarios require long chain-of-thought (CoT)[2] during translation. For example, in simple expressions, literal translation can meet most needs, and translation via long CoT may be unnecessary. Inappropriate scenarios might cause the overthinking issue (Chen et al., 2024).

ii) **A method to synthesize MT data with long thought**: Long thought SFT (supervised fine-tuning) data plays a vital role in simulating LLMs' long thought ability (Huang et al., 2024). Previous work pays much attention to how to synthesize long-thought data in math and coding tasks (Qin et al., 2024; Huang et al., 2024; Zhao et al., 2024).

For i), inspired by Van den Broeck (1981), a possible scenario is translating sentences with similes or metaphors, where literal translation often fails to convey the intended semantics. Given that, we decide to mine such sentences from literature books. The mining process uses an advanced large language model (LLM) to first judge `Q1`: *whether each literature sentence has any similes or metaphors*. If has, the LLM will be asked to literally translate the sentence to a target language, and give a final judgment on `Q2`: *whether literal translation is effective for native speakers of the target language to comprehend.* If the answers of `Q1` and `Q2` are "yes" and "no", respectively, the corresponding literature sentences will be reserved, and regarded as "suitable to translate via long thought".

For ii), after collecting the literal sentences with

[1]The codes, synthesized data and model checkpoints will be released upon publication.

[2]"long CoT" is equal to "long thought", and we alternatively use these two terms in this paper.

similes or metaphors, the next question is how to synthesize long thought MT samples. Previous work typically utilizes Monte Carlo Tree Search (MCTS) (Qin et al., 2024; Zhao et al., 2024; Zhang et al., 2024) or data distillation (Huang et al., 2024) (from existing O1-like models) to collect long thought SFT samples. Nevertheless, MCTS is typically used in math and coding tasks where multiple reasoning behaviors should be considered, and the method emphasizes complex reasoning that might not be efficient for machine translation. Besides, utilizing existing O1-like models for data distillation might (1) constrain the potential quality of the long-thought data; and (2) have a data gap in MT since current O1-like models are typically optimized toward math and coding tasks.

Therefore, we propose a multi-agent framework to synthesize MT data with long thought. In detail, there are three agents in the framework, *i.e.*, a translator, an advisor and an evaluator. The synthesis process is iterative, consisting of the following three steps during each iteration: (1) the translator generates a new translation conditioned on the previous step's translation and the corresponding refinement suggestions from the advisor; (2) the advisor evaluates the current translation and offers detailed feedback; (3) the evaluator assesses the current translation and gives an evaluation score using predefined scoring criteria. Once the translation score provided by the evaluator reaches a pre-defined threshold or the number of iterations reaches a maximum value, the iteration will stop. After that, the translation and suggestions in every step could form the long-thought MT samples. To improve the readability and fluency of the long-thought data, we employ GPT-4o (OpenAI, 2024a) to reformulate the long-thought content.

Based on the collected long-thought MT samples, we train our DRT-7B, DRT-8B and DRT-14B using the backbones of Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen2.5-14B-Instruct (Yang et al., 2024a), respectively. Experimental results on literature translation verify their effectiveness. In particular, DRT-14B outperforms QwQ-32B-preview and DeepSeek-R1-Distill-Qwen-32B in terms of BLEU, CometKiwi, CometScore and GPT-4 evaluations. Moreover, human evaluation and case study show the strong translation performance of DRT models.

Our main contributions are concluded as follows:

- We propose DRT aiming at building LLMs with long-thought machine translation ability. To achieve this, we mine literature sentences with similes or metaphors, and collect MT samples with long-thought processes.
- To synthesize the long-thought MT samples, we propose a multi-agent framework that involves a translator, an advisor and an evaluator. These three agents collaborate in an iterative manner to produce long thoughts during MT. Lastly, GPT-4o is used to further improve the quality of the synthesized long-thought MT samples.
- Experimental results on literature translation verify the effectiveness of our DRT. With the help of long thought, LLMs can learn to think during the machine translation.

2 DRT Data

We focus on English-to-Chinese translation[3], and we introduce how to collect the long-thought MT samples via three steps in this section: (1) collecting English sentences that tend to require long thoughts during translation (§ 2.1); (2) synthesizing the long-thought translation process for the collected sentences by a designed multi-agent framework (§ 2.2); (3) improving the readability and fluency of the long-thought content to form the final long-thought MT samples (§ 2.3). Next, we provide data statistics and data analyses of the collected data to give a deeper understanding (§ 2.4). Finally, we discuss the data quality (§ 2.5).

2.1 Literature Book Mining

Following Kryscinski et al. (2022), we leverage the literature books from the Project Gutenberg public-domain book repository[4], where the books are typically more than fifty years old and their copyrights have expired. About 400 English books are used to mine sentences with similes or metaphors.

First, we extract all sentences from these books, and filter out too short or too long sentences, *i.e.*, less than 10 words or more than 100 words, resulting in 577.6K literature sentences. Second, for each sentence, we use Qwen2.5-72B-Instruct (Yang et al., 2024a) to judge whether the sentence involves similes or metaphors, and discard the sentences that do not contain any ones. Third, for the remaining sentences, we let Qwen2.5-72B-Instruct literally translate them to Chinese, and then judge

[3] Although we focus on English-to-Chinese translation in this work, the methods we introduced can be trivially applied to other languages or translation directions.

[4] https://www.gutenberg.org/

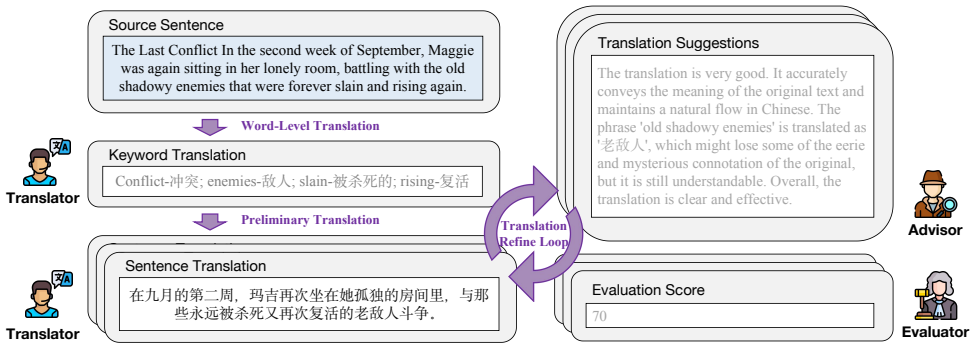

Figure 1: The illustration of the multi-agent framework to synthesize long-thought MT samples. (a) A translator iteratively produces translations under the suggestions provided by an advisor; (b) An advisor reviews the translation results and gives suggestions; (c) An evaluator assesses the translation results and gives an overall score to indicate the translation quality.

whether the translation satisfies native Chinese people. If the answer is negative, the corresponding sentence will be reserved, and regarded as "suitable to translate via long thought". For prompt details, please refer to Appendix A.1. Consequently, we collect 63K (out of 577.6K) literature sentences involving similes or metaphors whose literal translations have flaws, called *pre-collected sentences*.

2.2 Multi-Agent Framework

For each pre-collected sentence (denoted as s), we design a multi-agent framework to translate it via long thought. As shown in Figure 1, our multi-agent framework includes three agents: a translator, an advisor, and an evaluator, each of which use Qwen2.5-72B-Instruct as the backbone. The synthetic process is illustrated as follows:

(1) *Word-level Translation.* The translator first identifies the keywords that lie in the sentence, and then provides their translations under the consideration of the context. The keywords are denoted as $\mathcal{W}^{\text{src}} = \{w_1^{\text{src}}, w_2^{\text{src}}, ..., w_k^{\text{src}}\}$, where w_i^{src} indicates the i-th keyword in s, and k is the number of keywords. The translation of keywords is denoted as $\mathcal{W}^{\text{tgt}} = \{w_1^{\text{tgt}}, w_2^{\text{tgt}}, ..., w_k^{\text{tgt}}\}$. This step enables the model to identify potential challenges in translating the entire sentence by breaking it down into sub-problems (*i.e.*, word-level translation).

(2) *Preliminary Translation.* The translator then provides a preliminary sentence translation (t^0) conditioned on both the source sentence (s) and its keyword bilingual pairs ($\langle \mathcal{W}^{\text{src}}, \mathcal{W}^{\text{tgt}} \rangle$).

(3) *Translation Refine Loop.* In the refine loop, three agents work together to refine the translation iteratively. In each iteration step k (start from $k = 1$), the advisor first evaluates the translation in the previous step, *i.e.*, t^{k-1}, and provides detailed feedback f^{k-1} for polishing it. Then, the evaluator gives an overall score of t^{k-1} conditioned on both pre-defined scoring criteria and f^{k-1}, and the score is denoted as s^{k-1}. In the last of the iteration step, the translator takes its previous translation t^{k-1}, the corresponding feedback f^{k-1} and overall score s^{k-1} into account to provide a new translation t^k. The translation refine loop will stop when the overall score reaches a pre-defined threshold or the number of iteration steps meets the maximum. For prompt details of the translator, advisor and evaluator, please refer to Appendix A.2.

2.3 Long Thought Reformulation

After the multi-agent collaboration, we obtain a long thought process:

$$\begin{aligned} \mathcal{P}(s) : s &\Rightarrow \langle \mathcal{W}^{\text{src}}, \mathcal{W}^{\text{tgt}} \rangle \Rightarrow \langle t^0, f^0, s^0 \rangle \\ &\Rightarrow \langle t^1, f^1, s^1 \rangle \Rightarrow ... \Rightarrow \langle t^m, f^m, s^m \rangle \end{aligned} \tag{1}$$

where $\mathcal{P}(s)$ denotes the multi-agent thought process for s, and m is the number of iteration steps. To emphasize the valid thought process, translations without score change will be removed. That is, if s^i is equal to s^{i-1} ($i = 1, 2, ..., m$), we will discard $\langle t^i, f^i, s^i \rangle$ in $\mathcal{P}(s)$, resulting in:

$$\begin{aligned} \mathcal{P}'(s) : s &\Rightarrow \langle \mathcal{W}^{\text{src}}, \mathcal{W}^{\text{tgt}} \rangle \Rightarrow \langle t^0, f^0, s^0 \rangle \\ &\Rightarrow \langle t^{r_1}, f^{r_1}, s^{r_1} \rangle \Rightarrow ... \Rightarrow \langle t^{r_n}, f^{r_n}, s^{r_n} \rangle \end{aligned} \tag{2}$$

where $1 \leq r_1 < r_2 < ... < r_n \leq m$, and n is the number of remaining steps. If $n < 3$, we will discard the whole sample, *i.e.*, $\mathcal{P}(s)$.

For the remaining samples, we follow Qin et al. (2024), and leverage GPT-4o (OpenAI, 2024a) to modify and polish $\mathcal{P}'(s)$ into a self-reflection

[Source Sentence] Passing by the tower with her husband on the path to the gate she could feel the vibrant air humming round them from the louvred belfry in the circle of sound, and it matched the highly-charged mental atmosphere in which she was living.

[Thought During Translation] I start by identifying and translating the key terms from the English sentence: 'tower' as '塔', 'husband' as '丈夫', 'gate' as '大门', 'vibrant' as '充满活力的', 'humming' as '嗡嗡作响', 'louvred' as '有百叶窗的', 'belfry' as '钟楼', 'circle' as '圆圈', 'sound' as '声音', 'highly-charged' as '高度紧张的', 'mental' as '精神的', and 'atmosphere' as '氛围'.

With these translations in mind, I attempt my first translation: 经过塔时，她和丈夫走在通往大门的路上，能感受到有百叶窗的钟楼在声音的圆圈中发出的充满活力的空气在他们周围嗡嗡作响，这与她所处的高度紧张的精神氛围相匹配。

I reflect on this attempt and realize that while it is generally accurate, the phrasing could be improved for better fluency and clarity. The segment '有百叶窗的钟楼在声音的圆圈中发出的充满活力的空气' feels cumbersome and could be simplified. Additionally, '高度紧张的精神氛围' might not fully capture the nuance of 'highly-charged mental atmosphere.'

I revise my translation: 经过塔时，她和丈夫走在通往大门的路上，感受到从有百叶窗的钟楼传来的充满活力的空气在他们周围轻轻回响，这与她所处的高度活跃的精神氛围相契合。

This revision feels more fluent and natural, effectively conveying the original meaning. The phrase '轻轻回响' improves the flow and captures the sense of the air humming around them. The term '高度活跃的精神氛围' seems a better fit for 'highly-charged mental atmosphere,' making the translation more accurate and easier to understand for a Chinese reader.

I make another revision: 经过塔时，她和丈夫走在通往大门的路上，感受到从有百叶窗的钟楼传来的充满活力的气息在他们周围轻轻回荡，这与她内心的高度活跃氛围相契合。

This version is very smooth and natural, effectively conveying the original meaning. The use of '轻轻回荡' is particularly apt, capturing the sense of the air humming around them. The phrase '高度活跃的精神氛围' accurately reflects the 'highly-charged mental atmosphere,' making the translation both precise and fluent. There are no significant issues that would hinder understanding for a Chinese reader.

Finally, I decide on the final translation: 经过塔时，她和丈夫走在通往大门的路上，感受到从有百叶窗的钟楼传来的充满活力的气息在他们周围轻轻回荡，这与她内心的高度活跃氛围相映成趣。

Figure 2: An example of long thought synthesized by the designed multi-agent framework and GPT-4o reformulation.

	# Sample	Query	Thought	Output
o1-journey	327	41.53	486.05	3.41
Marco-O1 CoT data	10,000	52.73	673.98	52.73
DRT data (training)	19,264	37.25	527.64	44.67
DRT data (validation)	1,000	37.43	531.36	44.98
DRT data (testing)	2,000	37.19	525.44	44.70

Table 1: The number of samples and average token-level length of query, thought and output. "Query" and "Output" in DRT data mean the source sentences and the translated outputs, respectively.

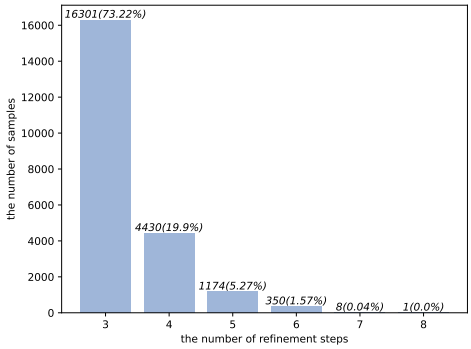

Figure 3: The distribution of the number of refinement steps in DRT data.

description (the used prompt is provided in Appendix A.3). Finally, we obtain 22,264 MT samples with long thought. Figure 2 gives an example sample to illustrate the synthetic results.

It is also worth noting that during the GPT-4o reformulation, we specify the translation with the highest score s^{r_j} as the final translation. Thus, the final translation is not necessarily the last one during refinement, *i.e.*, t^{r_n}.

2.4 Data Statistics and Data Analyses

We split the collected 22,264 samples into training, validation and testing sets with 19,264, 1,000 and 2,000 samples, respectively. Table 1 shows the data statistics of DRT data and previous O1-like data. For Marco-O1 CoT data (Zhao et al., 2024), since it is not fully released, we use its demo data to calculate the data statistics.[5] As we can see, the average number of tokens in our synthesized thought reaches 500+ tokens, showing the long thought process in our data.

Refine Loop Analyses. Figure 3 shows the number of refinement steps in the DRT data, which ranges from 3 to 8 steps. We can find that most samples (73.22%) involve 3 refinement steps, while only one sample involves 8 steps. Furthermore, to provide a deeper understanding of the refinement process, we calculate the average edit distance before and after each refinement step. Specifically, the first three refinement steps cause 21.44, 13.16 and 10.90 character-level edit distance. This observation is consistent with intuition. As the refinement progresses, the magnitude of the modification gradually decreases. To further understand the improvement brought by the translation refine loop, we calculate the average overall scores (provided

[5] https://github.com/AIDC-AI/Marco-o1

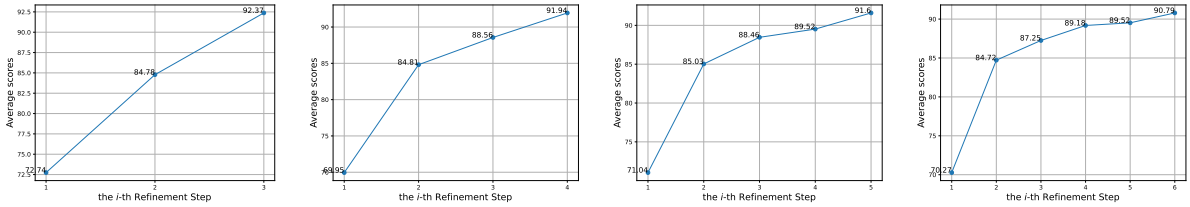

Figure 4: Trends in average scores (provided by the evaluator agent) over the refinement steps. The trends for samples with three, four, five, and six refinement steps are illustrated in (a), (b), (c), and (d), respectively.

Metric	ACC. (%)
CometKiwi	56.0
Evaluator Agent (Qwen2.5-72B-Instruct)	92.5
Evaluator Agent (GPT-4o)	93.5

Table 2: Accuracy of automatic metrics for translation quality estimation (ACC.: accuracy).

by the evaluator agent) along with each refinement step. As shown in Figure 4, as the number of refinement steps increases, the average score generally increases, demonstrating that the refine loop could iteratively increase the quality of translations.

2.5 Quality Analyses

The Effectiveness of the Evaluator Agent. Previous work has shown that the state-of-the-art LLMs can be used as evaluators for various text generation tasks (Kocmi and Federmann, 2023; Wang et al., 2023; Li et al., 2024). To figure out the effectiveness of our evaluator agent, we randomly select 200 source sentences from DRT data, and for each of them, we further select its two translations as well as scores (provided by the evaluator agent) during refinement. We next employ human annotators to compare the two translations of each source sentence, and judge which translation is better, or two translations are similar in quality (annotation details can be found in Appendix B). After obtaining the quality labels, we calculate the accuracy of the evaluator agent according to its evaluation score. For comparison, we also calculate the accuracy of CometKiwi (Rei et al., 2022) and GPT-4o evaluator agent. As shown in Table 2, our evaluator agent achieves a high accuracy (92.5%), demonstrating its effectiveness in evaluating literature translation quality. Besides, the widely-used CometKiwi metric only achieves 56.0% accuracy. Thought CometKiwi is powerful in the general domain (*e.g.*, news) (Kocmi and Federmann, 2023), its effectiveness in the literature domain is limited and unreliable, which is also pointed out by Karpinska and Iyyer (2023). Furthermore, the GPT-4o evaluator agent slightly outperforms the origin evaluator agent (with Qwen2.5-72B-Instruct backbone). Considering the tradeoff between cost and effectiveness, we finally decide to use Qwen2.5-72B-Instruct as our evaluator agent.

Translation Quality. Based on the effectiveness of the evaluator agent and the observation that evaluation scores of final translations typically reach 90.0 (c.f., Figure 4), we can ensure a high level of translation quality in the constructed data. According to the pre-defined scoring criteria of the evaluator agent (c.f., Appendix A.2), a score of 90.0 indicates excellent translations.

3 Experiments

3.1 Experimental Setups

Metrics. Following previous work, we adopt *"BLEU"* (Papineni et al., 2002), *"CometKiwi"* and *"CometScore"* (Rei et al., 2022) to evaluate the model translations. Among them, BLEU evaluates n-grams overlap between model translations and references, while CometScore evaluates the semantic similarity of model translations against references. CometKiwi uses a language model to judge whether a model translation conveys the semantics of the source sentence.

As pointed out by Karpinska and Iyyer (2023), BLEU and COMET may be ineffective for evaluating literature translation. Meanwhile, recent studies also show the strong ability of LLMs in NLP evaluation (Li et al., 2024). Therefore, we use evaluators implemented using GPT-4o in reference-based and reference-free styles, which we refer to as *"GRB"* and *"GRF"*, respectively. The evaluation prompts borrow from Kocmi and Federmann (2023), and are illustrated in Appendix C. Further-

Model	reference-free			reference-based		
	GEA	GRF	CometKiwi	GRB	BLEU	CometScore
Vanilla LLMs						
Llama-3.1-8B-Instruct	59.58	79.25	70.14	73.30	18.55	74.58
Qwen2.5-7B-Instruct	66.21	81.53	70.36	77.92	27.02	76.78
Qwen2.5-14B-Instruct	70.86	84.74	72.01	80.85	30.23	78.84
Marco-o1-7B	64.24	82.41	71.62	77.50	29.48	77.41
QwQ-32B-preview	75.50	86.31	71.48	83.08	27.46	78.68
DeepSeek-R1-Distill-Llama-8B	56.89	76.31	67.13	69.49	15.83	71.82
DeepSeek-R1-Distill-Qwen-7B	43.66	65.16	63.49	58.13	10.99	69.21
DeepSeek-R1-Distill-Qwen-14B	70.64	83.92	71.01	80.29	25.55	77.66
DeepSeek-R1-Distill-Qwen-32B	71.88	84.78	71.93	81.59	29.36	78.93
SFT LLMs (w/o CoT)						
Llama-3.1-8B-SFT	69.33	84.10	70.25	80.18	30.03	78.26
Qwen2.5-7B-SFT	72.29	85.06	71.03	81.72	35.44	80.10
Qwen2.5-14B-SFT	74.53	85.66	72.08	83.08	**37.63**	**80.82**
DRT						
DRT-8B (Backbone: Llama-3.1-8B-Instruct)	69.65†	84.49‡	70.85†	80.80†	32.67†	78.81†
DRT-7B (Backbone: Qwen2.5-7B-Instruct)	75.05†	85.57‡	71.78†	82.38†	35.54	80.19‡
DRT-14B (Backbone: Qwen2.5-14B-Instruct)	**77.41**†	**87.19**†	**72.11**	**83.20**‡	36.46	80.64

Table 3: Experimental results on literature translation. The **bold** and the underline denote the best and second-best performances, respectively. "†" and"‡" denote statistically significant better than the corresponding SFT LLMs (w/o CoT) with t-test $p < 0.01$ and 0.05, respectively.

more, as demonstrated in § 2.4, the GPT-4o evaluator agent achieves great accuracy in literature translation. We also leverage it as the evaluation metric in experiments, which is referred to as *"GEA"*. Since GRB, GRF and GEA need the API costs, we randomly select 400 samples to conduct evaluation.

Backbones. We adopt the following three LLMs as the backbones of our DRT: Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct (Yang et al., 2024b). All model checkpoints are publicly available.

For evaluation toolkits and the implementation details of all models, please refer to Appendix D.

3.2 Comparison Models

Vanilla LLMs. We leverage vanilla Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct (Yang et al., 2024a) as the comparison models. Besides, six O1-like LLMs are also conducted as baselines: Marco-o1-7B (Zhao et al., 2024), QwQ-32B-preview (Qwen, 2024), DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-14B and DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025).

SFT LLMs (w/o CoT). We also fine-tune LLMs with only paired sentences of DRT training data (without thought). This setting allows LLMs to learn the mapping from source literature sentences to the corresponding Chinese translations directly. We denote the fine-tuned LLMs as Llama-3.1-8B-SFT, Qwen2.5-7B-SFT and Qwen2.5-14B-SFT, serving as strong baselines in the experiments.

3.3 Main Results

Table 3 shows the experimental results, we analyze the performance from the following aspects:

SFT LLMs (w/o CoT) vs. Vanilla LLMs. After instruction tuning on the paired sentences of our training data, SFT LLMs (w/o CoT) significantly outperform the corresponding vanilla LLMs. For example, Llama-3.1-8B-SFT outperforms Llama-3.1-8B-Instruct by 9.75 GEA, 4.85 GRF and 6.88 GRB. Qwen2.5-7B-SFT outperforms Qwen2.5-7B-Instruct by 6.08 GEA, 3.53 GRF and 3.80 GRB. This finding demonstrates the effectiveness of our multi-agent framework and the quality of the synthesized translation. Please also note that the final translations are synthesized by Qwen2.5-72B-Instruct, indicating that we can leverage off-the-shelf *open-source* LLMs to collect high-quality literation translation data. And the data could help smaller LLMs (such as 7B and 14B ones) to boost their literature translation skills.

DRT vs. Vanilla LLMs. After fine-tuning on the long-thought MT training data, our DRT-series LLMs also significantly outperform the corresponding vanilla backbones. Particularly, DRT-14B outperforms QwQ-32B-preview and DeepSeek-R1-

Model	Flu.	Sem.	Lit.
Qwen2.5-14B-Instruct	-0.353	-0.363	-0.442
QwQ-32B-Preview	-0.063	0.022	-0.007
Qwen2.5-14B-SFT	0.103	0.108	0.087
DRT-14B	**0.313**	**0.233**	**0.362**

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

Distill-Qwen-32B in terms of all metrics, showing its effectiveness in literature MT.

DRT vs. SFT LLMs (w/o CoT). Using Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct as backbones, LLMs tuned with long thought achieve better performance than those tuned without long thought in terms of all metrics. For example, DRT-7B outperforms Qwen2.5-7B-SFT by 2.76 GEA, 0.51 GRF, 0.75 CometKiwi, 0.66 GRB, 0.10 BLEU and 0.09 CometScore. When using Qwen2.5-14B-Instruct as the backbone, we find that DRT-14B outperforms Qwen2.5-14B-SFT in terms of GEA, GRF, CometKiwi and GRB, but underperforms in terms of BLEU and CometScore. In detail, BLEU and CometScore evaluate the translations from the perspective of similarity between model translations and golden references. We conjecture that the higher BLEU and CometScore performance of Qwen2.5-14B-SFT is due to the model's ability to quickly learn domain-specific translations through tuning without long thoughts, allowing it to adapt to the literature translation more straightforwardly. However, training without long thoughts might lead the model to a sub-optimal solution, like learning shortcuts. When adopting evaluation metrics that are not significantly dependent on the golden references (*i.e.*, GEA, GRF, CometKiwi and GRB), DRT-14B shows its superior performance. Note that although GRB is a reference-based metric, it does not assess the model translations simply based on how similar they are to the golden references.

DRT vs. Commercial LLMs. To give a deeper understanding of our DRT models' performance, we further compare DRT models with GPT-4o (OpenAI, 2024a) and o1-preview (OpenAI, 2024b). The experimental results and corresponding analyses are provided in Appendix E.

3.4 Human Evaluation

We conduct human evaluation to further evaluate the performance of DRT-14B and strong baselines (Qwen2.5-14B-Instruct, QwQ-32B-Preview and Qwen2.5-14B-SFT). We randomly select 200 samples from our test set, and employ three human evaluators with high levels of fluency in English and Chinese to assess the generated translations from three aspects: fluency (Flu.), semantic accuracy (Sem.) and literary quality (Lit.). Following the Best-Worst Scaling method (Kiritchenko and Mohammad, 2017), evaluators are asked to select the best and the worst generated translation on 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. Thus, the final scores should range from -1 (worst) to 1 (best). As shown in Table 4, DRT-14B significantly outperforms these strong baselines, especially in the aspect of literary quality. These results demonstrate the superiority of our DRT models. The Fleiss' Kappa scores (Fleiss, 1971) of Flu., Sem. and Lit. are 0.75, 0.69 and 0.85, respectively, indicating a good inter-agreement among evaluators.

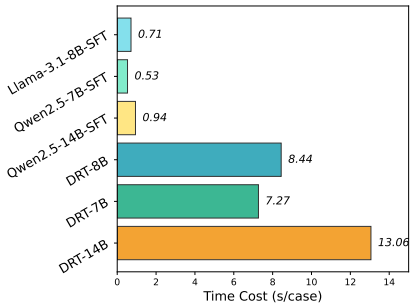

Figure 5: Time cost during inference on the testing set.

3.5 Inference Time Analysis

During evaluating LLMs' literature translation performance on our test set, we leverage vLLM to accelerate the model generation. A single NVIDIA A100 GPU (40G) is used to deploy each LLM. As shown in Figure 5, the average time costs of DRT models are significantly higher than LLMs (w/o CoT). This is because DRT models should first generate the long thought and then provide the final translation, thus needing more inference time ($\times$11.9~13.9). This also indicates that the O1-like LLMs may not be applicable to some scenarios with high real-time requirements.

3.6 Case Study

Table 5 provides some literature translation cases of Qwen2.5-14B-Instruct, QwQ-32B-Preview, Qwen2.5-14B-SFT and DRT-14B. With the help of long thought, the translations of DRT-14B align

Source Sentence	Qwen2.5-14B-Instruct	QwQ-32B-Preview	Qwen2.5-14B-SFT	DRT-14B
And if their words had taken another turn...if he himself had been less fastidious about intruding on another man's secrets...it was cruel to think how thin a film had shut out rescue from all this guilt and misery.	而如果他们的言辞转向了另一个方向.....如果他自己对侵扰他人秘密不那么介意.....想到正是那层薄薄的隔阂将救援拒之门外，使得这一切的罪恶和痛苦无法避免，真是残酷。	而且，假如他们的谈话走向不同.....假如他自己不对干涉他人私事如此谨慎.....那时他会意识到，只要一层薄薄的屏障就阻隔了救援，使得所有这些罪恶和痛苦无法解脱，这真是残忍。	如果他们的言辞稍有不同，若他不那么介意侵扰别人的隐私，想到那层薄纱如何隔绝了救援，真是残忍。	若他们的言辞稍作转变.....若他本人不那么介意侵入他人的心扉.....想到这层薄薄的隔阂竟将所有内疚与痛苦隔离于救援之外，实为残忍。
This cold officer upon a monument, who dropped epithets unconcernedly down, would be finer as a dead man, he thought.	他认为，这个站在纪念碑上的冷漠官员，若死了会更好，他不带任何感情地抛下了一些称呼。	这个冷冰冰的官员站在纪念碑上，毫不在意地抛下一些称号，他想，如果作为一个死人会更出色。	这位冷酷的官员站在纪念碑上，随意地抛出贬义词，他觉得，作为死人会更显尊贵。	纪念碑上的冷淡官员，漫不经心地吟咏那些修饰语，他心想，若化为亡者，或许更显尊贵。
I see on your cheek two tears which I know are hot as two sparks, and salt as two crystals of the sea.	我看见你脸颊上有两滴泪水，我知道它们像两簇火花一样滚烫，像两颗海水结晶一样咸。	我见你面颊两颗泪，温热如火花，咸涩如海晶。	我见你脸颊上两颗泪珠，热如火花，咸似海盐。	我看见你脸颊上的两滴眼泪，它们既如两簇火花般炽热，又似两粒海盐般咸涩。
The night wore out, and, as he stood upon the bridge listening to the water as it splashed the river-walls of the Island of Paris, where the picturesque confusion of houses and cathedral shone bright in the light of the moon, the day came coldly, looking like a dead face out of the sky.	夜深了，当他站在桥上，听着水花拍打巴黎岛河岸的声音，月光下房屋与教堂错落有致的景象闪耀着光芒时，冷冰冰的白天来临了，仿佛是从天空中露出的一张死人脸。	夜晚消逝了，当他站在桥上，聆听水流拍打着巴黎岛的河岸，那里的房屋与大教堂构成了一幅混乱而又美丽的图景，在月光下闪耀着明亮的光芒，而白天则冷酷地到来，仿佛是天空中一张死人脸。	夜幕渐渐消散，他站在桥上，聆听水花轻拍巴黎岛的河岸。月光下，岛上房屋与大教堂交织成一幅迷人的画卷，显得格外明亮。清晨悄然而至，宛如从天际浮现的一张冷峻的脸庞。	夜色渐渐消逝，他站在桥上，倾听河水轻拍巴黎城岛的河岸。月光下，房屋与大教堂错落交织，闪耀着迷人的光辉。白日冷冰冰地降临，宛如从天际浮现的一张死寂的脸。

Table 5: Case Studies of literature translation. Green indicates good translations, while red indicates bad ones.

more closely with the conventions of the Chinese language and exhibit a greater literary quality. In addition to DRT-14B, some translation snippets of other LLMs can also show a great performance (marked in green). This indicates that vanilla LLMs might have the capability to translate literature, and long thought could further activate this capability.

4 Related Work

O1-like LLMs. Recently, O1-like LLMs have shown great performance in reasoning tasks, especially math and coding tasks. After the emergency of OpenAI O1 model (OpenAI, 2024b), many efforts are given in reproducing OpenAI O1. For example, Qin et al. (2024) propose journey learning, a training paradigm, to encourage LLMs to learn not just shortcuts, but the complete exploration process. Huang et al. (2024) explore the data distillation from existing O1-like models, and show the effectiveness of data distillation. Zhang et al. (2024) leverage Monte Carlo Tree Search (MCTS) to synthesize reasoning-enhanced code data, and train O1-Coder. Marco-o1 (Zhao et al., 2024) is proposed to deal with open-ended text generation. More recently, DeepSeek-R1 (Guo et al., 2025) and Kimi K1.5 (Team et al., 2025) are proposed, and show their promising reasoning ability.

Literature Translation. Different from translating standard MT corpora (*e.g.*, news articles), translating literature books is more difficult since it often requires equivalence beyond the word level (Thai et al., 2022). Besides, it is also difficult to evaluate literature translation using automatic metrics, and previous literature translation work typically relies on human evaluation (Fonteyne et al., 2020; Karpinska and Iyyer, 2023). Due to its difficulty, early work is limited to small-scale attempts (Genzel et al., 2010; Jones and Irvine, 2013; Besacier and Schwartz, 2015; Toral et al., 2018). Recently, Karpinska and Iyyer (2023) utilize LLMs to perform literature translation, and show that discourse-level LLM translators achieve better performances compared with sentence-level approaches. Thai et al. (2022) introduce Par3 to benchmark LLMs' literature translation capability from non-English languages to English.

5 Conclusion

In this paper, we introduce DRT, an attempt to bring the success of long-thought reasoning to neural machine translation (MT). Specifically, we synthesize the machine translation long-thought samples by a designed multi-agent framework and GPT-4o reformulation. To collect the source sentences that are suitable for translation via long thought, we mine sentences with similes or metaphors from existing literature books. To synthesize the long thought machine translation process for these sentences, a translator, an advisor and an evaluator collaborate to translate the source sentence iteratively. Based on the synthesized data, we train DRT models. Extensive experiments on literature translation demonstrate the effectiveness of DRT models in terms of automatic evaluation. Case study and human evaluation further verify their superiority.

Limitations

While we show the effectiveness of long thought in MT, there are some limitations worth noting: (1) We focus on English-to-Chinese translation in this work, and future work could extend the data and the method to other translation directions. (2) There is still a lack of accurate automatic evaluation metrics for literary translation. Previous literature translation work typically relies on human evaluation (Fonteyne et al., 2020; Karpinska and Iyyer, 2023), and points out that BLEU and Comet might not be suitable for evaluating literature translation (Karpinska and Iyyer, 2023). This is because literary translations carry the responsibility of both semantic and critical interpretation, as they must address the challenge of achieving equivalence that often extends beyond the level of individual words (Thai et al., 2022).

Ethical Considerations

We discuss the main ethical considerations of DRT models as follows: (1) Copyright. We mine literature sentences from 400 English books provided by the Project Gutenberg public-domain book repository[6], where the books are typically more than fifty years old and their copyrights have expired. The book data also has been extracted and released by Kryscinski et al. (2022). Therefore, we can construct DRT data based on these books, and further release our synthesized data. (2) Licenses. We will release our model checkpoints and synthesized data under CC-BY-NC-SA 4.0 license.

References

Laurent Besacier and Lane Schwartz. 2015. Automated translation of a literary work: a pilot study. In *Fourth Workshop on Computational Linguistics for Literature-co-located with NAACL 2015*.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. 2024. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Josef Och. 2010. "poetic" statistical machine translation: rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166.

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

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*.

Ruth Jones and Ann Irvine. 2013. The (un) faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–101.

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.

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.

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

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[6] https://www.gutenberg.org/

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

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. 2024. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*.

Team Qwen. 2024. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*.

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.

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.

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

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Raymond Van den Broeck. 1981. The limits of translatability exemplified by metaphor translation. *Poetics today*, 2(4):73–87.

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.

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, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

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

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

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.

A Prompt in Data Synthesis

A.1 Prompts in Literature Book Mining

SYSTEM PROMPT:

You are assigned to translate an English literary work into Chinese. The text may include descriptions or expressions that embody English cultural nuances, which may not resonate with Chinese language habits. In such instances, a literal translation may not be appropriate; instead, these sentences should be paraphrased to convey their intended meaning effectively.

USER PROMPT:

The English sentence is provided as follows:
<english sentence>
{sentence}
</english sentence>

Please begin by assessing whether the English sentence contains any metaphors or similes. If there are none, respond with "no metaphors and no similes."

If the English sentence does contain metaphors or similes, provide a literal translation of them, and then evaluate whether the literal translation is appropriate and easy for Chinese natives to understand.

If it is suitable, format your response as follows (two lines):
"your literal translation for metaphors/similes here (in Chinese)"
"suitable"

If it is unsuitable, please provide the reason for the unsuitability. Format your response as follows (three lines):
"your literal translation for metaphors/similes here (in Chinese)"
"unsuitable"
"reason for unsuitability here (in Chinese)"

A.2 Prompts in Multi-Agent Framework

Translator Agent (Word-level translation)

Given an English sentence, identify the important words (usually nouns, verbs, technical terms, and named entities that require special attention in translation) and translate them into Chinese. Output the translations in JSON format, for example:

{"EnglishWord1": "ChineseTranslation", "EnglishWord2": "ChineseTranslation"}

The Chinese translations can be a single translation or multiple options as deemed appropriate.

Translator Agent (Preliminary translation)

SYSTEM PROMPT:

Given an English sentence and a JSON object containing potential translations of important keywords, produce a Chinese literal translation of the entire sentence. Please directly output the Chinese translation without any descriptions.

USER PROMPT:

<English Sentence>
{sentence}
</English Sentence>
<Potential Keyword Translation>
{keyword translation}
</Potential Keyword Translation>

Translator Agent (Refinement translation)

In the refine loop, the translator agent receives the feedback of the previous translation, and then provides a new translation. The prompt is a multi-turn dialogue between the translator and advisor, where the system prompt is the same as the preliminary translation.

Advisor Agent

Please rate the Chinese translation of the following English text and provide your comments and suggestions.

Evaluator Agent

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}

A.3 Prompts in Thought Reformulation

A student is engaged in the task of translating an English sentence into Chinese.

The English sentence is as follows:
<English Sentence>
{sentence}
</English Sentence>

This student constantly thinks about and optimizes his translation. The whole process is shown as follows:

<Translation Process>
{translation process}
</Translation Process>

Please polish the whole translation process into a long first-person self-reflection description (use the present tense).

The self-reflection should begin with selecting the keywords from the English sentence, translating the keywords, and then attempt to translate the whole sentence, and then think about whether the translation is good or not, and iteratively make translation attempts. Finally, make a final translation decision.

Output the self-reflection description directly without any additional descriptions or explanations. Each line in the self-reflection description can be regarded as a reasoning step to the translation.

B Details of Human Annotation

In Section 2.4, we employ human annotation to provide the quality comparison between two translations for a source sentence. Specifically, we employ three Chinese master students with high levels of fluency in both English and Chinese as our human annotators. For each sample, we give the source sentence and its two translation (without the scores provided by our evaluator agent) to all three annotators, and every annotator should provide one of the following judgments: (1) the first translation is better than the second one; (2) the second translation is better than the first one; (3) two translation are similar in quality. During annotation, we encourage the annotators to give differentiated judgments, *i.e.*, judgment (1) or judgment (2). If three annotators give the same judgment for a sample, the judgment will be regarded as the final label. Otherwise, its label will be decided by a group meeting of all three annotators and a senior data scientist.

C GPT-4o Evaluator

For GRB and GRF, we prompt GPT-4o (2024-08-06 version) as the MT evaluator in the reference-based and reference-free manners, respectively. The corresponding prompts borrow from Kocmi and Federmann (2023), and make some adaptions to literature translation.

GRB Prompt:

Score the following translation from English to Chinese with respect to the human reference on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect preservation of meaning, with faithfulness, expressiveness, and elegance".

English source: {src}
Chinese human reference: {ref}
Chinese translation: {hyp}

Score:

GRF Prompt:

Score the following translation from English to Chinese on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect preservation of meaning, with faithfulness, expressiveness, and elegance".

English source: {src}
Chinese translation: {hyp}

Score:

D Implementation Details.

Automatic Evaluation. To calculate CometKiwi and CometScore, we leverage the official codes[7] and the official models[8]. To calculate the BLEU score, we use the *sacrebleu* toolkit[9] to calculate the corpus-level BLEU.

Training Details. Llama-Factory (Zheng et al., 2024) is used to instruct-tune LLMs. All LLMs are tuned on 8×NVIDIA A100 GPUs (40G) with 1e-5 learning rate and 8 (8×1) batch size. We use the DeepSpeed ZeRO-3 optimization (Rasley et al., 2020). Following Qin et al. (2024), we set the number of training epochs to 3, and the training process costs 70 GPU hours and 124 GPU hours for 7B and 14B models, respectively.

Inference Details. When evaluating model performance on the test set, we use vLLM toolkit (Kwon et al., 2023) to accelerate the model generation. We

[7] https://github.com/Unbabel/COMET
[8] https://huggingface.co/Unbabel/wmt22-cometkiwi-da and https://huggingface.co/Unbabel/wmt22-comet-da
[9] https://github.com/mjpost/sacrebleu

Model	reference-free			reference-based		
	GEA	GRF	CometKiwi	GRB	BLEU	CometScore
Commercial LLMs						
GPT-4o	71.88	85.57	73.01	82.78	34.51	79.41
o1-preview	**78.01**	87.11	**73.70**	**83.86**	30.65	80.12
DRT						
DRT-8B (Backbone: Llama-3.1-8B-Instruct)	69.65	84.49	70.85	80.80	32.67	78.81
DRT-7B (Backbone: Qwen2.5-7B-Instruct)	75.05	85.57	71.78	82.38	35.54	80.19
DRT-14B (Backbone: Qwen2.5-14B-Instruct)	77.41	**87.19**	72.11	83.20	**36.46**	**80.64**

Table 6: Experimental results of comparing DRT with commercial LLMs. The **bold** and the underline denote the best and second-best performances, respectively.

use the sampling decoding strategy with 0.1 temperature, and set the repetition penalty to 1.05. For DeepSeek-R1 series (DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-14B and DeepSeek-R1-Distill-Qwen-32B), we follow the instruction[10] to enforce them to avoid blank thinking. All experimental results listed in this paper are the average of 3 runs.

E Comparison with Commercial LLMs

As shown in Table 6, DRT-14B achieves competitive results with o1-preview, showing its superiority. Additionally, we observe that o1-preview significantly outperforms GPT-4o in terms of GEA. This finding highlights the effectiveness of long thought in machine translation. When applied to appropriate translation contexts, long thought can further enhance the authenticity of translations.

[10] https://github.com/deepseek-ai/DeepSeek-R1