R1-T1: Fully Incentivizing Translation Capability in LLMs via Reasoning Learning

Minggui He[1], **Yilun Liu**[2†], **Shimin Tao**[2], **Hongyong Zeng**[3], **Jian Zhang**[1], **Yuanchang Luo**[2], **Li Zhang**[2], **Daimeng Wei**[2], **Weibin Meng**[2], **Osamu Yoshie**[1]

[1]Waseda University, Fukuoka, Japan
[2]Huawei Technologies Co., Ltd., China
[3]Xiamen University, Xiamen, China

Abstract

Despite recent breakthroughs in reasoning-enhanced large language models (LLMs), incorporating inference-time reasoning into application tasks such as machine translation (MT), where human translators naturally employ structured, multi-layered reasoning chain-of-thoughts (CoTs), is yet underexplored. Existing methods either design a fixed CoT tailored for a specific MT subtask (*e.g.*, literature translation), or rely on synthesizing CoTs unaligned with humans and supervised fine-tuning (SFT) prone to overfitting, limiting their adaptability to diverse translation scenarios. This paper introduces R1-Translator (R1-T1), a novel framework to achieve inference-time reasoning for general MT via reinforcement learning (RL) with human-aligned CoTs comprising six common patterns. Our approach pioneers three innovations: (1) verifying reasoning-based translation in various MT scenarios (*e.g.*, multilingual MT, domain MT) unseen from the training phase; (2) formalizing six expert-curated CoT templates that mirror hybrid human strategies like context-aware paraphrasing and round-trip translation; and (3) enabling more flexible CoTs through an RL stage after cold-start. Both human and automatic evaluation results indicate a steady translation quality improvement in a total of 10+ languages and 40+ translation directions on Flores-101 test set and four domain-specific MT tasks, especially on the languages unseen from training.

[1]† Corresponding author: liuyilun3@huawei.com.

1 Introduction

Recent advancements in large language models (LLMs) have brought prominent paradigms of R1-like models (DeepSeek-AI, 2025), which leverage chain-of-thought (CoT) reasoning at inference time (Wei et al., 2022) for tasks like mathematical proofs and code generation, and employ reinforcement learning (RL) to incentivize autonomous reasoning behaviors without supervised fine-tuning (SFT) (Shao et al., 2024). For instance, DeepSeek-R1 (DeepSeek-AI, 2025) exemplifies how extended CoT reasoning through decomposed intermediate steps can tackle complex problem-solving. However, the application of extended CoT reasoning to machine translation (MT)—a task demanding both linguistic fidelity and contextual reasoning—remains underexplored.

While LLMs have shown promise in MT, prior work primarily focused on direct SFT on parallel corpus (Kocmi et al., 2024a) or retrieval-augmented methods (Wang et al., 2025b). Recently, Wang et al. (2025a) proposed DRT (*Deep Reasoning Translation*) and introduced long CoT to enhance Chinese-English literary translation, revealing that extended reasoning steps improve handling of culturally nuanced phrases and metaphors. Yet, this approach narrowly targeted literature, overlooking the broader spectrum of translation scenarios where professional translators rigorously apply structured reasoning—such as decomposing idioms, resolving contextual conflicts, and adopting techniques (*e.g.*, round-trip translation) to refine outputs (Ordudari, 2007; Nida, 1964; Somers, 2005). Nida (1964) also stated that a technical translator should make "a thorough study of the source language text

Transactions of the Association for Computational Linguistics, vol. 14, pp. 2103–2119, 2026. https://doi.org/10.1162/TACL.a.793
Action Editor: Derek F. Wong. Submission batch: 3/2026; Revision batch: 4/2026; Published 8/2026.

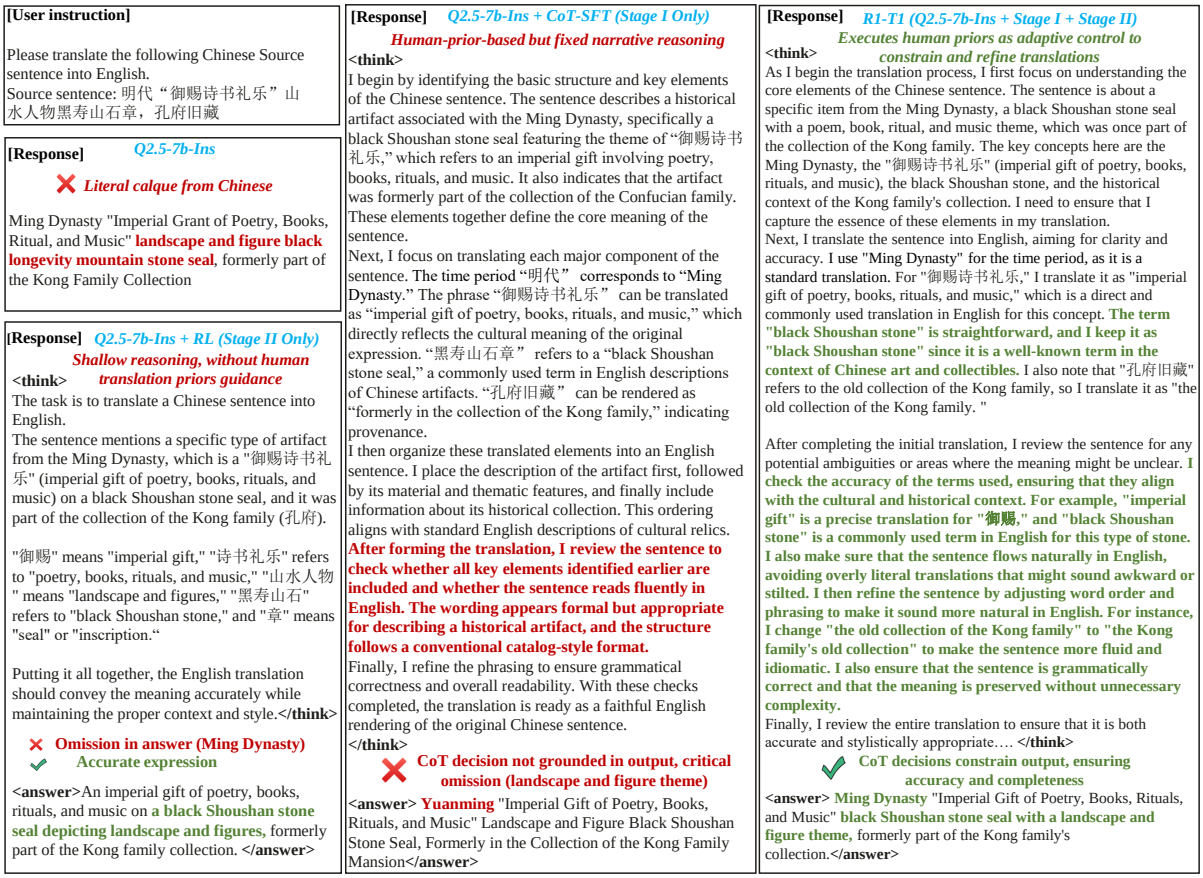

Figure 1: Case study illustrating how different reasoning paradigms affect translation quality. The base model (Q2.5-7b-Ins) relies on surface-level literal mapping, leading to calque errors. CoT-SFT (Stage I) reproduces human-prior-based reasoning as a fixed narrative, but fails to ground identified decisions in the final output. Q2.5-7b-Ins + RL (Stage II only) induces shallow reasoning without structured human translation priors, resulting in incomplete translations. In contrast, R1-T1 executes human CoT strategies not as a static template, but as a flexible guideline, enabling reasoning decisions to actively constrain and refine translation outputs.

before making attempts to translate it, ”suggesting a deep thinking process in human translation. Current MT systems lack mechanisms to reflect these human-like, complex CoT processes, limiting their adaptability to diverse translation challenges.

Moreover, existing approaches for reasoning-based translation utilized purely synthesized CoTs distilled from LLMs (*e.g.*, DRT (Wang et al., 2025a)) or designed a single fixed procedure (*e.g.*, TASTE (Wang et al., 2024)), which may be sub-optimal and unaligned with the hybrid reasoning strategies employed by human experts. Also, their reliance on SFT with pre-defined CoTs can easily cause overfitting (Chu et al., 2025a), restricting generalization ability to flexibly organize CoTs for broader translation domains and tasks. To address these, we propose **R1-Translator (R1-T1)**, a framework that fully incentivizes reasoning-based translation through three innovations:

(1) Reasoning for general and broader MT. Beyond literature translation, we extended reasoning-based MT to general multilingual MT and MT tasks in diverse domains such as culture, common sense and terminology constraint, while expanding language support to 10+ languages for multilingual general MT, including low-resource languages. This aligns with our core assumption, drawing from human translation experiences, that reasoning is beneficial for general and broader MT.

(2) Human-aligned translation CoTs through expert strategies. Professional translators often interleave complex reasoning procedure such as lexical disambiguation, context-aware paraphrasing, and iterative self-correction. With the help of human translators, we formalized these into six distinct CoT templates mimicking common thinking patterns of human translators and curated a training dataset with hybrid translation CoT trajectories.

Unlike single-type CoT in existing approaches, our approach mirrors the multi-layered reasoning observed in human workflows for generalization.

(3) Flexible and dynamic CoTs by RL incentivization. To avoid merely "memorizing" predefined CoT trajectories (Chu et al., 2025a) and increase generalizability of learned human translation knowledge, we designed an RL process with specially designed rewards for MT after cold-start, enabling LLMs to dynamically refine its reasoning strategies to earn the best translation rewards. As shown in Fig. 1, consequently, the model is equipped with a more flexible reasoning ability while still guided under our curated human strategies (*i.e.*, checking the specific terms for accuracy according to learned human strategies), instead of merely repeating the definition of the strategy mechanically observed for the CoT-SFT-only (*i.e.*, SFT using pre-defined CoTs) model without RL.

To realize this vision, we introduce the concept of *translation reasoning learning*, training R1-T1 models through two interconnected stages: (1) SFT on hybrid CoT trajectories curated with professional translators, and (2) a RL-based refinement to flexibly calibrate its reasoning paths. Stage one serves as a cold start, injecting initial human reasoning guidance on translations to the model and facilitating RL convergence. Stage two ensures the flexible application of human prior knowledge without overfitting. Our contributions are:

- We verified the benefits of reasoning for various MT scenarios, achieving superior translation quality in 10+ languages and multiple sub-tasks.
- We identified six reasoning strategies for translation commonly adopted by human translators and validate the effectiveness of incorporating CoT trajectories built with these human prior knowledge into reasoning-based MT.
- We pioneered the successful application of RL on MT to improve the generalizability of reasoning ability, by designing a novel reward policy for MT, which significantly improves the translation quality compared with plain SFT strategy.

In addition, we open-source R1-T1's datasets and code, facilitating future research on MT. [2]

[2]Code, dataset, reasoning templates are available at https://github.com/superboom/R1-T1.

2 Related Work

2.1 Reasoning-based LLMs

The pioneering advancements in reasoning-based LLMs, such as OpenAI's O1 (Jaech et al., 2024) and DeepSeek-R1 (DeepSeek-AI, 2025), have excelled in many tasks and attracted significant research attentions. While earlier explorations focused on using inference-time reasoning for solving complex tasks such as math and coding (Qin et al., 2024; Zhang et al., 2024a), there is a trending belief towards utilizing reasoning-based LLMs for general AI tasks. Zhao et al. (2024) expanded reasoning-based LLMs to open-ended text generation, generalizing them to broader domains where clear standards and quantified rewards are absent. Shen et al. (2025) proposed VLM-R1, a stable and generalizable R1-style LLM for vision-language tasks. In addition, reasoning-based LLMs applied for financial tasks (Chu et al., 2025b) and adapted to multilingual reasoning (Ko et al., 2025).

Compared with existing endeavors on reasoning-based LLMs, our work focuses on introducing long CoTs into the field of general MT, thereby developing a reasoning-based LLM for MT.

2.2 LLMs for Machine Translation

Since the appearance of ChatGPT (Ouyang et al., 2022), endeavors on applying LLMs for MT continuously emerge. Early studies focused on directly prompting LLMs to translate via prompt strategies (Jiao et al., 2023b, 2024; Peng et al., 2023). Other works fine-tuned open-source LLMs using parallel corpus (Xu et al., 2024; Wu et al., 2024; Zhang et al., 2023b). ParroT (Jiao et al., 2023a) further advanced this line by reformulating MT into an instruction-following paradigm with explicit *Hint*-based constraints, enabling more controllable translation via instruction tuning. Techniques were further proposed to enhance the quality of LLMs for the tasks of MT, such as continuous pre-training (Boughorbel et al., 2024; Fujii et al., 2020), mixture-of-expert modules (Xu et al., 2025; Zhu et al., 2025) and multi-task training (Wang et al., 2024; Ul Hassan et al., 2024).

Recently, MT with CoT methodology has drawn the attention from researchers. Feng et al. (2025) introduced an API-based self-correcting framework for LLMs, where LLMs autonomously call external evaluation models and refine translation hypothesis based on the evaluation result. Wang et al. (2024) designed a multi-task training phase and a

multi-stage inference phase for MT, guiding the LLM to first conduct translation task, then a evaluation task and a revision task. DRT (Wang et al., 2025a) merged this procedure into an inference-time CoT, utilized multi-agent mechanism to synthesize such long CoTs and examined quality in English-Chinese literature translation.

Compared with existing MT approaches, our R1-T1 incorporates six types of human-aligned CoTs for translation, possesses the ability to flexibly applying these CoTs via RL, and is designed for broader MT.

3 Methodology

The overview of R1-T1 is shown in Fig. 2. To incentivize reasoning capabilities in LLM-based translators, we first construct the MT Reasoning Dataset, which introduces human-aligned reasoning trajectories into a seed parallel corpora via CoT translation template. R1-T1 is then trained on the MT Reasoning Dataset through the two stages in translation reasoning learning. Section 3.1 provides a detailed description of the construction process. In Section 3.2, we outline the component implementation of *translation reasoning learning*.

3.1 Construction of MT Reasoning Dataset

With the goal of enabling LLMs to mimic the human expert reasoning process in translation, we propose a framework that explicitly incorporates human reasoning strategies into the MT seed dataset via a *Human-Guided AI Synthesis* mechanism.

As shown in Fig. 2, we identify six core reasoning strategies grounded in professional translator practice. These strategies are operationalized into structured reasoning modules as reusable CoT templates, where each strategy is expressed as explicit, step-by-step instructions to make translators' tacit expertise reproducible. By instantiating these templates with our carefully prepared parallel corpus, the model can learn from reasoning processes in real translation examples. Notably, the instantiation process aims to provide explicit and human-aligned insights into the model's reasoning pathways, which helps accelerate convergence of further learning.

Reasoning Strategies in Human Translation. We first elicited a broad set of translation reasoning strategies through semi-structured interviews with 8 professional translators, whose backgrounds are introduced in Appendix A. To standardize terminology and reasoning scope, we cross-referenced the elicited patterns with prior work in translation studies and NLP literature, and validated the reformulations with the translators.

With the pool of reformulated strategies, we conducted three rounds of discussion to consolidate candidate strategies by (i) merging overlapping scopes that were not functionally distinguishable in practice, (ii) refining strategy definitions, and (iii) discarding strategies that either did not play a distinct functional role in the translation process or could not be reliably abstracted into stable, step-by-step CoT procedures. This process resulted in the final six reasoning strategies used in this work, which capture the commonly recurring, functionally distinct patterns observed across translators.

The six strategies are detailed as follows:

(1) *Hierarchical Translation:* Reflecting human translators' practice of segmenting complex texts into manageable components. By identifying essential elements first, the model translates each segment with greater contextual awareness, reducing errors from overlooked structural dependencies.

(2) *Triangulation Translation:* Also known as relay translation (Ringmar, 2012), it employs an intermediate (pivot) language to bridge linguistic and cultural gaps. When direct translation is difficult, pivoting enhances grammatical accuracy, vocabulary appropriateness, and cultural nuance.

(3) *Round-trip translation:* Translating a target sentence back to the original language to detect and correct translation errors or inconsistencies (Somers, 2005), ensuring improved fluency and fidelity.

(4) *Context-aware Translation:* Emphasizing the broader context (e.g., paragraph or document) in translation decisions (Cabezas-García, 2023; Almanna, 2015), this strategy effectively preserves idiomatic expressions, cultural nuances, and domain-specific terminologies.

(5) *Translation Explanation:* Providing rationales for translation choices when multiple interpretations are possible (Bühler, 2002), enhancing transparency and allowing for more informed and justified translation outcomes.

(6) *Structural Transformation:* Systematically adjusting sentence structures to adhere to target language syntactic norms, essential for translations between structurally distinct languages such as English to Japanese or German. This ensures gram-

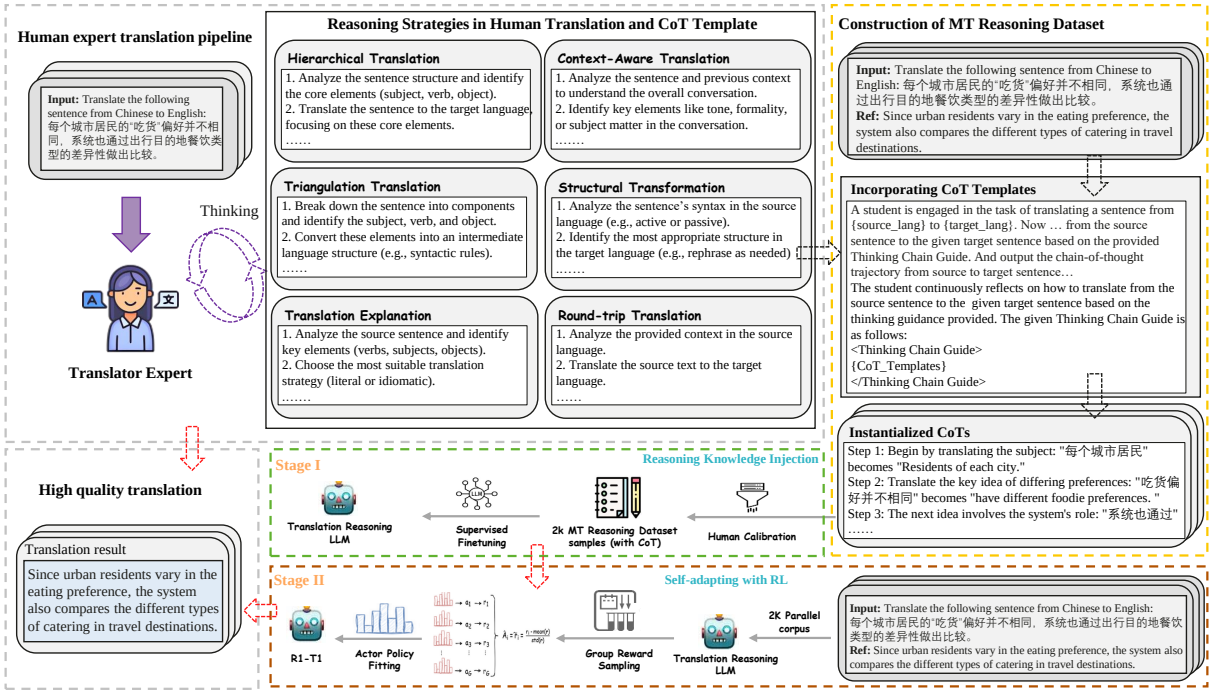

Figure 2: Illustration on the construction of MT reasoning dataset and training of R1-T1.

matical correctness and syntactic appropriateness.

In real-world practice, these six strategies form a cohesive workflow for professional translators. For example, Liu, a senior translator, encountered a complex sentence in a cultural policy text: "面对错综复杂的非遗保护现状，我们不应一刀切，而要因地制宜。" (*Given the complex landscape of intangible heritage protection, we should avoid one-size-fits-all measures and adopt locally tailored approaches.*) She first decomposed the sentence into its core issue (heritage protection) and policy stance to clarify the underlying logic (Hierarchical Translation). Recognizing the UNESCO policy context, she determined that the idiom "一刀切" should be rendered in a formal rather than colloquial manner (Context-aware Translation). To translate the culturally specific expression "因地制宜," she consulted equivalent expressions across languages (e.g., the French *sur mesure*) as a conceptual pivot, leading to the formulation "tailored approaches" (Triangulation Translation). Among multiple candidates, she justified the standardized term "Intangible Cultural Heritage (ICH)" to ensure terminological consistency (Translation Explanation). She then reorganized the sentence structure to align with English policy discourse conventions (Structural Transformation), and finally verified the draft by mentally mapping it back to the source meaning to correct subtle modality shifts (Round-trip translation). This example illustrates how the six strategies function as complementary reasoning lenses in professional translation practice.

Construction of CoT Templates from Human Experience. Drawing upon the six core translation reasoning strategies, we abstracted structured CoT templates directly from human translator's workflows, formulating each strategy into detailed, step-by-step instructions to enable LLMs to consistently instantiate explicit reasoning trajectories.

For example, the *Hierarchical Translation* template explicitly guides the LLM through several structured stages: initially analyzing the source sentence to identify core linguistic elements (subject, verb, object), then performing a preliminary translation specifically focusing on these core elements to ensure foundational accuracy. Subsequently, the model reviews the translation to identify areas needing refinement, such as word choice, grammatical correctness, or contextual coherence, and iteratively improves the translation's fluency and readability. Finally, it validates that the refined translation accurately preserves the original meaning and aligns well with the broader textual context. We systematically developed similar step-by-step instructions for each of the other five templates and iteratively validated them with the translators.

Each template has been meticulously crafted and calibrated by experienced human translators to authentically reflect human translation workflows, enhancing transparency, and facilitating clear and effective instantiation by LLMs (all templates are available in our repository[2]).

Parallel Corpus Collection. Our seed dataset consists of approximately 2k parallel translation pairs sampled from open source real-world MT datasets. [3][4][5] Data include translations at sentence and paragraph level, with token lengths ranging from 10 to 1200, yielding a length distribution comparable to established MT benchmarks (Bañón et al., 2020; Goyal et al., 2022). Token length is computed by tokenizing the source text using the Qwen2.5-7B tokenizer (Yang et al., 2024) and counting the length. Specifically, sources include (1) news data from the WMT corpus (Kocmi et al., 2024a), (2) formal texts from the UN corpus (Kocmi et al., 2024a), and (3) domain-specific corpora containing terminology-rich translations. For language coverage, we focus on six main languages: Russian (ru), French (fr), German (de), Japanese (ja), Chinese (zh), and English (en), selected for their large speaker populations and cross-linguistic diversity (Bojar et al., 2018). From the above sources, we identify 20 translation directions within these six languages. For each direction, parallel pairs are randomly sampled from the source corpora, with a light constraint on token-length coverage to ensure diversity across short and long inputs, resulting in approximately 2k translation pairs after filtering. Table 1 displays the statistics.

Incorporating CoT Templates into Parallel Corpus. With calibrated templates, we instantiate translation CoTs for each sample in our seed dataset using an annotation executor under a systematically controlled procedure, resulting in 2k parallel pairs augmented with reasoning trajectories. In pilot annotation trials, we worked with human translators to evaluate several candidate LLMs as annotation executors, including Qwen2.5-72B-Instruct, Qwen2.5-Max (Yang et al., 2024), and GPT-4o (Hurst et al., 2024), and adopted GPT-4o as the annotation executor because it adhered most reliably to our handcrafted reasoning templates and required the least human revision effort.

Language Pair	Avg. Tokens per Sample	# Examples	Domain
en→zh, ru, fr, de, ja	398.5	493	Terminology, News
zh→en, ru, fr, ja	116.2	400	UN Corpus, News literature
fr→zh, ru, de, en	320.4	399	UN Corpus, News
ja→zh, en	280.8	200	Terminology, News
de→en, fr	404.6	193	Terminology, News
ru→en, fr, zh	229.6	296	UN Corpus

Table 1: Statistics for seed parallel corpus. **en, zh, ru, fr, de, ja** are the abbreviations for English, Chinese, Russian, French, German and Japanese, respectively. **Avg. Tokens per Sample** refers to the average number of tokens in the source language input.

Given a source sentence and its reference translation, the executor is prompted to (i) extract key semantic anchors from the source (e.g., core concepts, entities, and critical relations), (ii) select suitable template(s) from our six human-defined strategies conditioned on these anchors, and (iii) instantiate the selected template(s) into an explicit, step-by-step reasoning trajectory. This forms a complete training sample consisting of a *source sentence*, a *reference translation*, and an *explicit reasoning trace*.

All reasoning traces are instantiated in English, following evidence that multilingual LLMs exhibit latent reasoning centered on English and benefit from pivot-language transfer, which stabilizes intermediate reasoning and facilitates cross-lingual generalization (Schut et al., 2025; Zhang et al., 2024b). To ensure data quality and human faithfulness, human translators manually inspected all generated samples to verify source/reference faithfulness and correct template adherence before inclusion in training.

3.2 Translation Reasoning Learning

Stage I: SFT with Reasoning Dataset as Cold Start In the process of *translation reasoning learning*, we initiate the training process with a cold start phase, leveraging the MT Reasoning Dataset in Section 3.1 to fine-tune the model. This approach draws inspiration from Deepseek-R1 training strategies (DeepSeek-AI, 2025). During this stage, the model is trained to jointly predict the explicit reasoning process and the final translation output, conditioned on the source sentence. The cold start phase establishes a foundational

[3] https://machinetranslate.org/wmt

[4] https://www.un.org/dgacm/zh/content/uncorpus/ch

[5] https://huggingface.co/datasets/joefox/newstest-2017-2019-ru_zh/tree/main

layer of reasoning-aligned generation, in which the model learns to follow human-annotated step-by-step reasoning trajectories while producing faithful translations. More importantly, human professional knowledge of translation is injected into the model via paired reasoning–translation examples, which facilitates convergence and stabilizes subsequent reinforcement learning.

Stage II: Self-adapting with RL This part introduces a self-adapting RL mechanism that allows the model to progressively refine its reasoning skills over time after the cold start phase. Notably, only the 2k training parallel corpus in Table 1 without CoT are used in this phase. This process facilitates a spontaneous improvement on its translation strategies by a rewarding signal on long-term feedback (*i.e.*, overall translation quality of its hypothesis). The reward calculation and learning strategy for RL are as follows:

RL Reward Modeling: For the RL stage, we have designed a reward modeling system consisting of two distinct types of rewards: *(a) Format Reward:* This reward encourages the model to adhere to a structured response format. We use regular-expression extraction to ensure that the reasoning process is placed within the "<think></think>" tags and the final translation is enclosed within the "<answer></answer>" tags. *(b) Answer Reward:* This reward evaluates the translation quality of the model output. We adopt COMET-20 (Rei et al., 2020) as the reward model, as it produces scores with wider separation across quality levels compared to newer variants such as COMET-22 (Rei et al., 2022) whose scores are bounded within [0, 1], which providing more informative reward distinctions for group-based advantage estimation in our RL optimization.

However, several challenges arise during the design of this system: **a**. In practical applications, relatively large negative rewards cause the agent to become overly conservative, potentially trapping it in local optima (Schulman et al., 2017). **b**. When the scale of translation reward is much lower than the format reward, the model prioritizes format correctness over translation quality, leading to suboptimal behavior. To address these challenges, we first set the lower bound of the reward functions to 0 to prevent negative rewards. To further balance the rewards, we set the upper limit of the format reward to 0.2, as suggested by Kocmi et al. (2024b) for cases of low translation quality. This ensures that the format reward does not overly interfere with the translation quality, allowing the model to focus on both translation quality and formatting. Thus, the reward system r is computed as:

Hyper-Parameter		SFT	RL
Optimizer	Batch size	64	16
	Epochs	2	2
	Learning rate	1×10^{-4}	4×10^{-7}
	Warmup ratio	0.05	0.1
Token Limits	Max prompt length	2300	2300
	Max response length	4000	4000
Sampling	Temperature	–	0.8
	Num. Rollout	–	12
RL Policy	KL coefficient (β)	–	0.001
	PPO mini-batch size	–	256
	PPO micro-batch size	–	128

Table 2: Hyperparameter settings.

$$r_{\text{form}} = \begin{cases} 0.2 & \text{if the format is correct} \\ 0 & \text{if the format is incorrect} \end{cases} \quad (1)$$

$$r_{\text{ans}} = \begin{cases} 0 & \text{if } x \leq 0 \\ Com(trans, ref) & \text{if } x > 0 \end{cases} \quad (2)$$

$$r = \begin{cases} 0 & \text{if } r_{\text{form}} = 0 \\ r_{\text{ans}} + r_{\text{form}} & \text{if } r_{\text{form}} > 0 \end{cases} \quad (3)$$

Here, Com refers to the COMET-20 score, x refers to Com(trans, ref), trans is the generated translation and ref is the reference translation. r_{form} and r_{ans} denote the Format Reward and Answer Reward, respectively.

RL Strategy: For the RL phase, we employ the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024), which was first publicly proposed to optimize reasoning LLMs using complex rewards without a critic model, employing group-based advantage estimation.

4 Experiment

4.1 Experiment Setting

In this section, we provide an overall evaluation for R1-T1, covering general MT, domain-specific MT, ablation study, and the case study of self-evolution translation CoT during RL.

Implementation Details we selected Qwen2.5-7B-Instruct as a backbone model due to its superior multilingual performance among open source models with fewer than 10 billion parameters (Yang et al., 2024). This choice helps minimize potential negative impacts from insufficient base model capabilities. The MT Reasoning Dataset was randomly split into a training set and a validation set at a ratio of 9:1. For GRPO implementation, we develop it based on VeRL Platform[6]. Both SFT and GRPO are conducted with full-parameter training, and all detailed hyperparameters are listed in Table 2.

Dataset	ru	fr	de	ja	zh	en
Flores101	0.142	0.137	0.097	0.113	0.110	0.097
Common sense	-	-	-	-	0.071	0.089
Terminology	-	-	0.106	-	-	0.108
Culture	-	0.135	0.106	-	0.102	0.123
Literature	-	-	-	-	0.102	0.105

Table 3: Token overlap between test set and our training set. The overlap is measured by Jaccard overlap, which is the intersection-over-union of unique tokens between train and test corpus.

Evaluation Dataset We evaluated our model under general purpose and domain-specific benchmarks. For general multilingual translation, we employ *Flores-101* (Goyal et al., 2022), which offers extensive coverage of diverse languages across a wide range of topics, making it suitable for evaluating general MT capabilities. In addition, we added four domain-specific benchmarks to evaluate robustness under challenging translation scenarios. *CommonsenseMT* (He et al., 2020) evaluates models' ability to resolve lexical and syntactic ambiguities using commonsense knowledge in Chinese-English translation. *RTT* (Zhang et al., 2023a) assesses MT robustness to terminology constraints using an English-German benchmark with multi-term inputs that challenge constraint fidelity. *CultureMT* (Yao et al., 2024) measures cultural translation adequacy via a multilingual benchmark annotated with culture-specific items requiring pragmatic adaptation. *DRT Literature* (Wang et al., 2025a) offers literary texts rich in figurative language, testing models' ability to preserve metaphors, similes, and stylistic fidelity. To avoid data leakage, Table 3 summarizes the token-level overlap between these evaluation sets and our training data, which produced a similarity score of less than 0.15.

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

Metric We use both automatic and manual evaluation. For automatic evaluation, we employ xCOMET-XL (Guerreiro et al., 2024) and MetricX-24-XXL (Juraska et al., 2023) to measure the quality of translation hypotheses given references. xCOMET (ranging from 0 to 1) is an explainable learned metric that provides a sentence-level quality score and has shown strong correlation with human judgments. MetricX-24 (ranging from 0 to 25) is a learned regression-based metric trained to predict translation quality scores aligned with MQM-style(Freitag et al., 2021) ratings. For human evaluation, we cooperated with 30 professional translators covering 12 languages across EN/ZH $\rightarrow$ XX and XX $\rightarrow$ EN/ZH directions. Each translation was independently scored by three annotators as a standalone, anonymous item, blind to system identities and without access to other systems' outputs for the same source. Annotators assigned scores following the evaluation guidelines and provided brief comments on observed issues. The evaluation assesses Accuracy, Fluency, and Direct Assessment (DA). Accuracy and Fluency are rated using a 5-point Likert scale (1: severe breakdown; 5: no critical errors), while DA is assigned on a 0–100 scale. Sentence-level averages are reported. Detailed guidelines are provided in Appendix B.

4.2 Main Results

Our main experiment focuses on model performance of multilingual MT, investigating how well its translation quality in both trained and unseen languages against existing approaches.

Baselines We compare our R1-T1 against several representative models with multilingual MT capability, including (1) General-Purpose LLMs: *Qwen2.5-7B-base*, *Qwen2.5-7B-Instruct*, *Qwen2.5-14B-Instruct*, *Llama3.1-8B*; (2) Reasoning LLMs: *DeepSeek-R1-Distill-Qwen-7B*, *DeepSeek-R1-Distill-Qwen-7B-Multilingual* [7], and Marco-o1 (Zhao et al., 2024); and (3) Translation models: *ParroT-7B* (Jiao et al., 2023a), *TEaR* (Feng et al., 2025), and *NLLB* (NLLB Team et al., 2022). For fair comparison, TEaR is instantiated with *Qwen2.5-7B-Instruct* as its underlying LLM, following the original setup. All baselines are publicly available, and DRT is only compared in the domain *en2zh* translation setting due to its design and lingual limitations.

[7] https://huggingface.co/lightblue/DeepSeek-R1-Distill-Qwen-7B-Multilingual

Models	xx⇔en (MetricX (↓) & Xcomet (↑))					xx⇔zh (MetricX (↓) & Xcomet (↑))					Avg.
	zh	ja	ru	fr	de	en	ja	ru	fr	de	
General-purpose LLMs											
Q2.5[a]-7b	1.81 / 0.892	4.23 / 0.657	3.25 / 0.673	2.41 / 0.938	2.35 / 0.709	1.81 / 0.892	3.30 / 0.613	2.90 / 0.834	2.49 / 0.820	2.48 / 0.644	2.80 / 0.753
Q2.5-7b-Ins[b]	2.36 / 0.811	4.16 / 0.753	3.73 / 0.775	2.69 / 0.823	2.35 / 0.855	2.36 / 0.811	3.81 / 0.725	4.13 / 0.703	3.39 / 0.738	3.21 / 0.774	3.31 / 0.773
Q2.5[a]-14b-Ins[b]	1.72 / 0.904	3.36 / 0.882	3.16 / 0.886	2.25 / 0.940	**1.99** / 0.954	1.72 / 0.904	3.31 / 0.841	3.32 / 0.788	2.28 / 0.828	2.24 / 0.880	2.63 / 0.878
la3.1[c]-8b-Ins[b]	7.90 / 0.701	3.81 / 0.872	8.95 / 0.713	15.71 / 0.434	10.00 / 0.663	7.90 / 0.701	4.79 / 0.784	4.74 / 0.794	7.46 / 0.700	4.69 / 0.845	7.56 / 0.723
Reasoning LLMs											
DS-R1-D[d]	12.54 / 0.657	19.37 / 0.645	21.08 / 0.650	20.42 / 0.765	18.87 / 0.792	12.54 / 0.657	12.41 / 0.627	14.38 / 0.607	13.49 / 0.625	15.02 / 0.687	16.40 / 0.673
DS-R1-DM[e]	6.48 / 0.847	15.67 / 0.594	19.15 / 0.651	17.51 / 0.765	16.53 / 0.805	6.48 / 0.847	7.54 / 0.626	6.97 / 0.614	6.07 / 0.622	11.03 / 0.695	11.88 / 0.691
Marco-o1	1.80 / 0.903	3.75 / 0.870	3.19 / 0.915	2.62 / 0.929	2.40 / 0.952	1.80 / 0.903	3.56 / 0.827	2.98 / 0.828	2.52 / 0.817	2.45 / 0.878	2.81 / 0.880
Translation Models											
ParroT-7b	5.20 / 0.730	7.21 / 0.634	7.12 / 0.786	3.67 / 0.890	3.06 / 0.939	5.20 / 0.730	5.34 / 0.781	4.26 / 0.842	4.43 / 0.819	3.24 / 0.879	4.84 / 0.811
NLLB	6.74 / 0.714	8.30 / 0.680	7.88 / 0.759	7.85 / 0.738	7.03 / 0.804	6.74 / 0.714	8.23 / 0.598	9.16 / 0.584	8.83 / 0.538	8.39 / 0.634	8.05 / 0.672
TEaR	1.79 / 0.902	3.68 / 0.868	3.60 / 0.887	2.92 / 0.917	2.57 / 0.948	1.79 / 0.902	3.64 / 0.816	3.65 / 0.794	2.79 / 0.796	2.78 / 0.862	3.05 / 0.866
R1-T1	**1.55**† **/ 0.915**†	**3.15**† **/ 0.902**†	**2.73**† **/ 0.929**†	**2.20 / 0.944**	**1.99 / 0.962**†	**1.55**† **/ 0.915**†	**3.11**† **/ 0.847**‡	**2.44**† **/ 0.850**‡	**2.12**† **/ 0.838**†	**2.03**† **/ 0.895**†	**2.37 / 0.898**

Table 4: R1-T1's translation quality on **Trained** languages in Flores-101. **xx⇔en** means the average quality of **xx** to English and English to **xx** (the same for **xx⇔zh**). **Q2.5** means Qwen2.5. **Ins** means Instruction. **DS-R1-D(M)** means DeepSeek-R1-Distill-Qwen-7B(-Multilingual). **Bold** and underline denote the best and second-best scores in each translation direction, respectively. ↑ / ↓ indicate that higher / lower values are better. † and ‡ denote statistically significant improvements over the second-best baseline with $p < 0.01$ and $p < 0.05$, respectively. zh, ja, ru, fr, de are Chinese, Japanese, Russian, French, and German, respectively.

Models	xx⇔en (MetricX / XCOMET)					xx⇔zh (MetricX / XCOMET)					Avg.
	th	nl	vi	tr	cs	th	nl	vi	tr	cs	
General-purpose LLMs											
Q2.5[a]-7b	3.22 / 0.863	3.33 / 0.918	2.79 / 0.906	7.25 / 0.716	6.40 / 0.855	3.38 / 0.776	2.97 / 0.819	2.63 / 0.834	7.58 / 0.637	6.21 / 0.746	4.58 / 0.807
Q2.5-7b-Ins[b]	4.60 / 0.710	3.68 / 0.796	3.73 / 0.767	8.37 / 0.639	6.62 / 0.710	4.67 / 0.650	4.39 / 0.699	3.99 / 0.705	8.59 / 0.553	7.25 / 0.626	5.59 / 0.686
Q2.5[a]-14b-Ins[b]	3.23 / 0.853	3.62 / 0.905	3.44 / 0.853	6.98 / 0.774	**5.06 / 0.887**	3.39 / 0.768	2.79 / 0.830	4.40 / 0.755	6.60 / 0.682	**4.86 / 0.784**	4.44 / 0.809
la3.1[c]-8b-Ins[b]	11.27 / 0.643	10.94 / 0.622	9.27 / 0.715	16.54 / 0.426	9.54 / 0.736	7.10 / 0.681	4.25 / 0.825	5.43 / 0.779	8.46 / 0.689	5.36 / 0.769	8.82 / 0.689
Reasoning LLMs											
DS-R1-D[d]	15.03 / 0.546	12.11 / 0.661	13.75 / 0.501	16.44 / 0.447	15.27 / 0.519	10.86 / 0.533	10.63 / 0.558	11.92 / 0.517	15.29 / 0.435	14.55 / 0.455	13.58 / 0.517
DS-R1-DM[e]	14.06 / 0.542	11.25 / 0.684	12.42 / 0.555	17.78 / 0.393	14.81 / 0.538	13.11 / 0.449	10.55 / 0.588	13.60 / 0.575	15.82 / 0.409	13.34 / 0.474	13.67 / 0.521
Marco-o1	3.62 / 0.852	3.46 / 0.917	2.87 / 0.898	7.94 / 0.762	6.68 / 0.841	3.60 / 0.767	3.07 / 0.817	3.00 / 0.819	7.03 / 0.695	5.89 / 0.750	4.72 / 0.812
Translation Models											
ParroT-7B	18.43 / 0.283	4.37 / 0.881	16.27 / 0.387	18.11 / 0.367	7.90 / 0.795	12.82 / 0.522	4.14 / 0.767	10.83 / 0.584	10.56 / 0.599	5.21 / 0.772	10.86 / 0.596
NLLB	11.39 / 0.560	7.46 / 0.764	7.11 / 0.738	9.81 / 0.692	9.66 / 0.711	10.86 / 0.485	8.72 / 0.572	7.62 / 0.617	10.95 / 0.531	11.94 / 0.520	9.55 / 0.619
TEaR	4.24 / 0.805	4.20 / 0.877	3.65 / 0.848	9.15 / 0.702	6.98 / 0.824	4.10 / 0.737	3.66 / 0.787	3.60 / 0.787	7.87 / 0.636	7.05 / 0.691	5.45 / 0.769
R1-T1	**3.11**‡ **/ 0.869**‡	**3.03**† **/ 0.928**†	**2.47**† **/ 0.912**†	**6.82**† **/ 0.796**†	6.13 / 0.859	**3.00**† **/ 0.793**†	**2.70**‡ **/ 0.837**	**2.41**† **/ 0.841**‡	**5.96**† **/ 0.715**†	5.58 / 0.767	**4.12 / 0.832**

Table 5: R1-T1's translation quality on **Unseen** languages in Flores-101. **th**, **nl**, **vi**, **tr**, and **cs** refer to Thai, Dutch, Vietnamese, Turkish and Czech, respectively. Definition of notations are the same as Table 4.

Result of Trained Languages As shown in Table 4, R1-T1 consistently achieves the best overall translation quality in all translation directions involving languages with training data in Flores-101, with the lowest average MetricX score of 2.37 and the highest average XCOMET score of 0.898. Compared to general-purpose LLMs such as the Qwen2.5 series and LLaMA3.1-8B-Instruct, R1-T1 demonstrates clear and statistically significant improvements in most translation directions, particularly for xx⇔zh. Reasoning-oriented models like DS-R1-D series show less stable translation quality, with large variances across directions, suggesting that their reasoning mechanisms are not as effective in multilingual translation. Marco-o1, on the other hand, performs more competitively in several directions, though it still trails R1-T1 overall. Translation-specific models like TEaR perform well, but it is a multi-agent translation system, which differentiates its translation quality from the single-agent R1-T1. ParroT-7B and NLLB also show competitive translation quality in certain directions, but still trail R1-T1. This highlights the advantages of explicitly integrating reasoning supervision into translation-oriented training.

Result of Unseen Languages As shown in Table 5, the advantage of R1-T1 is more significant in unseen language translation (*i.e.*, achieving 16 best out of the 20 testing terms), highlighting its robust generalization. Due to the relative limited resources of the five languages in Table 5 (*e.g.*, not covered by the WMT news corpus), most existing methods struggle to achieve a steady translation quality compared with the common languages in Table 4 (*e.g.*, average XCOMET drops sharply from 0.773 to 0.686 for Qwen2.5-7B-Ins). In contrast, R1-T1, despite not training on these lan-

Models	Cmn.	Literature		Term
	zh2en	en2zh	zh2en	en2de
Q[a]2.5-7b-Ins[b]	2.31	4.39	4.19	4.20
Q[a]2.5-14b-Ins[b]	1.75	4.01	3.33	2.84
la3.1[c]-8b-Ins[b]	8.58	8.27	9.63	2.49
Marco-o1	1.71	5.43	3.37	4.06
ParroT-7B	4.41	10.66	7.33	**2.34**
TEaR	2.00	4.39	3.46	4.21
DRT-o1-7b	–	**3.56**	–	–
R1-T1	**1.62**‡	3.95	**3.02**†	3.08

Models	Culture						avg
	zh2en	en2zh	es2en	en2es	fr2en	en2fr	
Q[a]2.5-7b-Ins[b]	4.14	4.16	4.30	3.73	5.98	4.68	4.50
Q[a]2.5-14b-Ins[b]	3.65	3.67	4.21	3.28	5.90	4.16	4.15
la3.1[c]-8b-Ins[b]	11.40	8.96	16.50	10.43	17.58	12.13	12.83
Marco-o1	3.83	3.88	4.69	3.95	6.45	5.12	4.65
ParroT-7B	8.21	8.50	4.35	4.15	6.08	5.32	6.10
TEaR	3.97	3.97	4.51	5.08	6.11	6.20	4.97
DRT-o1-7b	–	3.46	–	–	–	–	–
R1-T1	**3.47**†	**3.32**†	**4.13**‡	**3.19**‡	**5.79**†	**4.09**‡	**4.00**

Table 6: R1-T1's translation quality on Domain-specific MT evaluated by MetricX (↓). Cmn. means Common Sense. Each cell reports MetricX. Definition of notations are the same as Table 4.

guages, benefiting from RL-enhanced reasoning, consistently achieves superior COMET and MetricX scores. This confirms that the step-by-step reasoning paradigm can decompose complex and unfamiliar problems into solvable steps, enhancing generalizability.

4.3 Evaluation on Domain-specific MT

As stated in Section 4.1, we evaluate R1-T1 across four challenging domains: *commonsense*, *terminology*, *culture*, and *literature*. The results, shown in Table 6, indicate that R1-T1 outperforms all baseline models in average translation quality across these domains. Note that DRT-o1-7b achieves higher scores in the literature for en2zh, possibly due to its extensive exposure to en2zh literary data during training. LLaMA3.1-8B-Instruct and ParroT-7B perform strongly in terminology for en2de, but struggles in other domains (*e.g.*, 5.43 and 10.66 in literature for en2zh). These findings align with our central hypothesis that RL-enhanced reasoning training enables better generalization to challenging domain tasks. For common sense or terminology scenarios, reasoning encourages the model to better preserve critical expressions and maintain contextual accuracy. In cultural or literary, CoT guidance helps the model disambiguate nuanced expressions and interpret figurative symbols within the appropriate situational context.

Models	F101	Cmn.	Lit	Term	Culture
Q2.5-7b	4.84	2.31	4.29	4.20	4.50
Q2.5-7b + SFT	4.19	1.97	3.82	3.53	4.10
Q2.5-7b + RL (w/ reasoning)	3.54	1.71	3.64	3.75	4.28
Q2.5-7b + RL (w/o reasoning)	3.66	1.83	3.73	4.03	4.36
Q2.5-7b (w/ human CoT) + SFT	4.53	1.95	4.72	6.16	5.15
Q2.5-7b (w/o human CoT) + SFT + RL	4.97	2.36	4.71	5.98	4.91
Q2.5-7b + RL (w/ comet22)	6.19	2.53	6.97	7.81	7.60
R1-T1					
Q2.5-7b (w/ human CoT) + SFT + RL	**3.20**†	**1.67**‡	**3.48**‡	**3.08**†	**4.00**

Table 7: Ablation study evaluated by MetricX (↓). F101 = Flores-101; CmnS = Common Sense; Lit = Literature. Notation definitions are consistent with Table 4.

4.4 Ablation Study

Ablation on the RL incentivization for MT We trained and compared models with identical training settings except for distinct training paradigms and evaluated their translation quality on various benchmarks. Specifically, Table 7 reports eight settings in total. Settings (1)–(6) are used to analyze the effect of different RL training paradigms, the comparison between (6) and (7) examines the contribution of human expert-aligned CoT data, and setting (8) investigates the effect of reward model choice: (1) *Q2.5-7b*: Original Qwen2.5-7b-instruct without additional fine-tuning. (2) *Q2.5-7b + SFT*: SFT applied directly using the 2k parallel translation corpora without CoT. (3) *Q2.5-7b + RL (w/ reasoning)*: Directly conducting stage II of the *translation reasoning learning* (i.e., RL with reasoning tokens). (4) *Q2.5-7b + RL (w/o reasoning)*: Conducting RL optimization without introducing explicit reasoning tokens. (5) *Q2.5-7b (w/ human CoT) + SFT*: Conducting only stage I (i.e., SFT with human CoT). (6) *Q2.5-7b (w/ human CoT) + SFT + RL*: The full R1-T1 training pipeline with both stages. (7) *Q2.5-7b (w/o human CoT) + SFT + RL*: A counterpart initialized with CoT trajectories directly distilled from GPT-4o without our human-designed reasoning templates. (8) *Q2.5-7b + RL (w/ comet22)*: Replacing COMET-20 with COMET-22 as the reward model for (3).

As demonstrated in Table 7, the following key findings emerge clearly: a. Direct RL optimization with reasoning (setting (3)) consistently outperforms the original model (setting (1)) and pure SFT baselines (setting (2)) on most benchmarks, demonstrating that reward-driven optimization effectively improves translation quality. b. Directly applying human CoT supervision through SFT alone (setting (4)) leads to inferior results across benchmarks. This suggests that rigid reasoning templates, when

used only as supervised targets, may restrict model flexibility without adaptive reward optimization. c. In contrast, when human CoT is used as cold-start initialization for subsequent RL (setting (6)), the full R1-T1 pipeline achieves the best overall translation quality, showing that structured human priors and adaptive RL training are complementary. d. Replacing COMET-20 with COMET-22 as the reward model (setting (7)) leads to substantial degradation, even underperforming the base model. This suggests that COMET-22 provides less stable advantage estimates in our GRPO setting, thereby preventing effective reward optimization.

Ablation on human expert-aligned CoT data We further examine the contribution of the proposed MT Reasoning Dataset by comparing settings (5) and (6). Under identical RL settings, replacing our expert-aligned CoT with reasoning traces generated from GPT-4o leads to consistent quality degradation, confirming that the gains stem from the structured human translation priors in our reasoning templates rather than the executor model's own capability.

4.5 Human Evaluation

As automatic metrics may not fully capture nuanced translation defects (Freitag et al., 2021; Agrawal et al., 2024), we conduct human evaluation on 570 uniformly sampled instances from the five test domains, following the rigorous evaluation criteria outlined in Appendix B. Table 8 shows that R1-T1 achieves the best DA and Accuracy scores among the three models across all domains. In particular, the improvements are statistically significant on Flores-101, Common Sense, and Term in both DA and Accuracy, and on Term for Fluency. This pattern aligns with our objective of improving translation via reasoning: the human-prior reasoning strategies and their RL refinement are particularly effective for multilingual and terminology-related domain MT tasks. The summarized translator comments in Table 9 further illustrate recurring advantages over both baselines in terminology usage, structural organization, and source faithfulness. Improvements in Literature and Culture are more moderate. Table 9 indicates that these domains involve knowledge-intensive challenges, such as culturally grounded word-choice issues in culture-related cases. Since such errors require background knowledge and culturally grounded interpretation beyond structural

Models	Flores-101			Common sense			Literature		
	Flu	Acc	DA	Flu	Acc	DA	Flu	Acc	DA
Q2.5-7b + SFT	4.34	4.21	86.16	**4.50**	4.05	82.17	4.37	3.93	77.36
Marco-o1	4.19	4.06	82.11	4.43	4.08	81.61	4.11	3.75	74.13
R1-T1	**4.36**	**4.33**‡	**87.41**‡	4.46	**4.47**‡	**87.57**‡	**4.48**	**4.03**	**79.02**

Models	Term			Culture		
	Flu	Acc	DA	Flu	Acc	DA
Q2.5-7b + SFT	4.03	4.50	88.16	**4.58**	4.17	83.92
Marco-o1	3.88	4.41	86.56	4.25	4.02	79.67
R1-T1	**4.68**†	**4.80**‡	**93.17**†	4.49	**4.20**	**84.25**

Table 8: Human evaluation results across five domains. Fluency (Flu.), Accuracy (Acc.), and Direct Assessment (DA), respectively. Higher is better. Overall inter-annotator agreement ICC(A,3)=0.796.

reasoning and contextual coherence, the advantages brought by our reasoning-oriented training are less directly reflected in these domains, leading to smaller translation quality gaps. Overall, the human evaluation confirms reliable gains of R1-T1 in multilingual and terminology-related domain MT tasks.

4.6 Reasoning Impact on Translation Analysis

To provide a more intuitive understanding of how reasoning affects translation quality, as shown in Fig. 1, we compare different reasoning paradigms to highlight their impact on translation quality. The base model (Q2.5-7b-Ins) relies on literal translation, leading to frequent calque errors. CoT-SFT (Stage I) introduces fixed human-prior templates, reducing errors but lacking flexibility, as its reasoning does not adapt to the context, leading to occasional omissions. In contrast, RL-only (Stage II) introduces shallow reasoning through reinforcement learning, improving the base model by better capturing context but still lacking structured decision-making, limiting consistency. R1-T1 combines the strengths of both, using adaptive CoT templates that refine reasoning and actively constrain translation output, ensuring both flexibility and accuracy. This progression from rigid reasoning templates to adaptive control highlights how R1-T1 optimizes translation by dynamically adjusting reasoning based on context, resulting in more accurate and contextually aware translations. We further compare R1-T1 with two reasoning-oriented baselines in Fig. 3. DeepSeek-R1-Distill-Qwen-7B exhibits term mis-identification and hallucinated content, while Marco-o1 produces verbose but weakly grounded reasoning that leaves key ambiguities unresolved. R1-T1 remains source-grounded through-

Language	EN/ZH → XX	XX → EN/ZH
Arabic	R1-T1 generally shows accurate wording and clear expression. Q2.5+SFT has occasional imprecise wording. Marco exhibits grammar issues, mistranslations, and excessive literal translation in some cases.	R1-T1 produces accurate terminology with concise expression. Q2.5+SFT shows lexical imprecision in some cases. Marco has occasional natural phrasing but errors in names and currency terms.
German	R1-T1 uses accurate wording with fluent output, though both R1-T1 and Q2.5+SFT show occasional vocabulary and grammar errors. Marco mistranslates domain terms, invents non-existent compounds, and has grammar errors.	R1-T1 overall shows fluent logical flow, especially in conjunction usage and restructuring. Q2.5+SFT has occasional mistranslations. Marco shows named-entity errors and occasionally produces untranslated content.
Spanish	R1-T1 generally provides accurate and natural translations with occasional minor errors. Q2.5+SFT shows redundancy, omissions, and mistranslations. Marco has serious entity/geography term errors and non-standard expressions.	R1-T1 generally achieves accuracy and fluency with occasional term or logic errors. Q2.5+SFT has terminology errors and unnatural phrasing. Marco shows unclear references and less fluent expression.
French	R1-T1 is generally natural and source-faithful, though sometimes has term issues. Q2.5+SFT shows additions, omissions, and mistranslations. Marco has unnatural style and entity mistranslations.	R1-T1 generally achieves accuracy and fluency. Q2.5+SFT reads fluently in some subsets. Marco underperforms both with literal renderings and omissions.
Russian	R1-T1 generally shows accurate terminology and faithful structure. Q2.5+SFT tends to add unsupported content. Marco is faithful but occasionally uses imprecise word choices.	R1-T1 and Q2.5+SFT show similar quality. R1-T1 is faithful to source; Q2.5+SFT reads naturally. Marco performs better in this direction but shows named-entity errors.
Chinese–English	R1-T1: accuracy and faithfulness. Q2.5+SFT: omissions and additions. Marco: source closeness but lexical errors.	R1-T1 and Q2.5+SFT: similar adequacy; both show named-entity errors. Marco: natural but frequent named-entity errors.

Table 9: Summary of human evaluator comments on R1-T1, Q2.5-7B+SFT and Marco-o1 across representative translation directions and languages. Each entry is summarized by the authors from per-example annotator comments collected during independent evaluation. EN/ZH → XX means translation directions from English or Chinese to target languages, and vice versa.

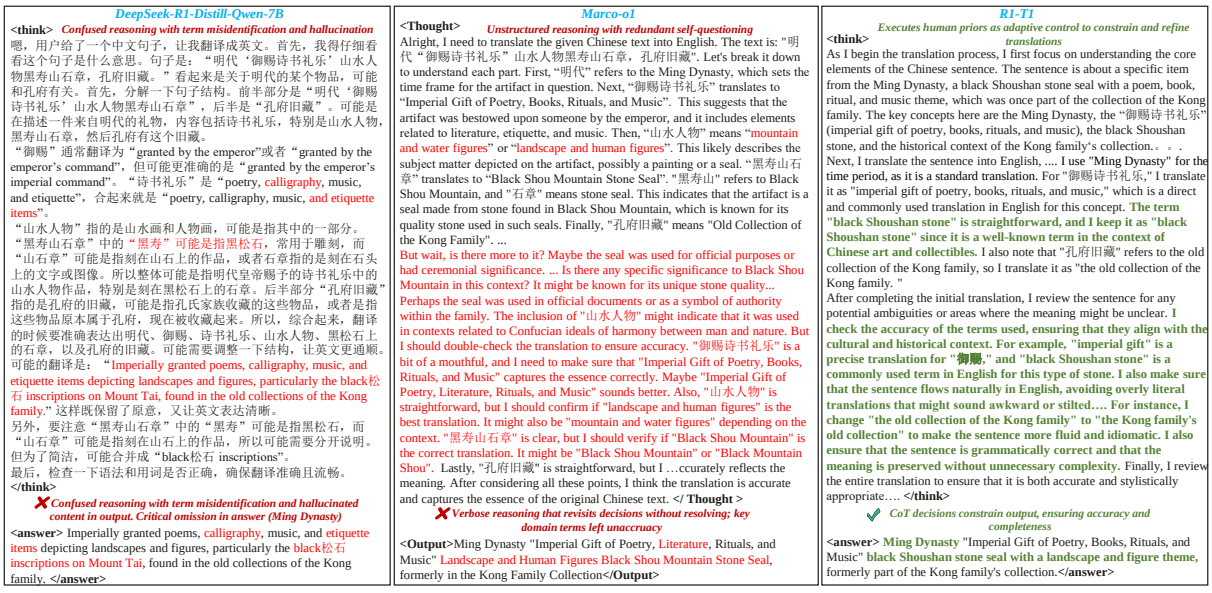

Figure 3: Case study comparing two reasoning baselines (DeepSeek-R1-Distill-Qwen-7B and Marco-o1) and R1-T1 on the same translation example.

out, highlighting a clearer link between reasoning grounding and translation quality.

5 Conclusion

In this paper, we revealed the benefits of performing general MT tasks with inference-time reasoning, by combining human-aligned translation CoTs and RL strategies tailored for MT tasks. The success of R1-T1, our proposed MT framework, on MT tasks across multiple domains and languages reflects the necessity and wide existence of the reasoning processes in human translations, which are ignored by existing methods. Moreover, the self-adapted paradigm of CoTs in translation reasoning learning provide R1-T1 with high flexibility in generalizing its translation ability to unseen languages, domains and tasks, as indicated by the improved translation quality in these unseen scenarios. Future work include incorporating more languages and CoTs in training and testing in more scenarios.

Acknowledgments

We thank the action editor and anonymous reviewers for their careful review and responsible suggestions, which are very helpful for improving our work.

References

Sweta Agrawal, António Farinhas, Ricardo Rei, and André F. T. Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14438–14457. Association for Computational Linguistics.

Ali Almanna. 2015. *Contextualizing translation theories: Aspects of Arabic–English interlingual communication*. Cambridge Scholars Publishing.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, et al. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Sabri Boughorbel, Md Rizwan Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 73–88, Bangkok, Thailand. Association for Computational Linguistics.

Axel Bühler. 2002. Translation as interpretation. *Translation studies: Perspectives on an emerging discipline*, pages 56–74.

Melania Cabezas-García. 2023. The use of context in multiword-term translation. *Perspectives*, 31(2):365–382.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025a. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning*, pages 10818–10838. PMLR.

Xu Chu, Zhijie Tan, Hanlin Xue, Guanyu Wang, Tong Mo, and Weiping Li. 2025b. Domaino1s: Guiding LLM reasoning for explainable answers in high-stakes domains. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3330–3348. Association for Computational Linguistics.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948v1*. Version referenced: v1.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. TEaR: Improving LLM-based machine translation with systematic self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Atsushi Fujii, Masao Utiyama, and Eiichiro Sumita. 2020. Continual learning for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

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

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

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276v1*. Version referenced: v1.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720v1*. Version referenced: v1.

Hui Jiao, Bei Peng, Lu Zong, Xiaojun Zhang, and Xinwei Li. 2024. Gradable chatgpt translation evaluation. *Procesamiento del Lenguaje Natural*, 72:73–85.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023b. Is ChatGPT a good translator? yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745v4*. Version referenced: v4.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767. Association for Computational Linguistics.

Hyunwoo Ko, Guijin Son, and Dasol Choi. 2025. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 78–95, Suzhuo, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings*

of the Ninth Conference on Machine Translation, pages 1–46. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024b. Preliminary WMT24 ranking of general MT systems and LLMs. *arXiv preprint arXiv:2407.19884v1*. Version referenced: v1.

Eugene Albert Nida. 1964. *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672v3*. Version referenced: v3.

Mahmoud Ordudari. 2007. Translation procedures, strategies and methods. *Translation journal*, 11(3):8.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. o1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982v1*. Version referenced: v1.

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

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.

Martin Ringmar. 2012. Relay translation. *Handbook of translation studies*, 3:141–144.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347v2*. Version referenced: v2.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

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 arXiv:2402.03300v3*. Version referenced: v3.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615v2*. Version referenced: v2.

Harold Somers. 2005. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 127–133, Sydney, Australia.

Muhammad Naeem Ul Hassan, Zhengtao Yu, Jian Wang, Ying Li, Shengxiang Gao, Shuwan Yang, and Cunli Mao. 2024. Lkmt: Linguistics knowledge-driven multi-task neural machine translation for urdu and english. *Computers, Materials & Continua*, 81(1).

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. DRT: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Yingxue Zhang, and Jie Zhou. 2025b. Retrieval-augmented machine translation with unstructured knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5858–5871. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024. TasTe: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6144–6158, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468v4*. Version referenced: v4.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115v2*. Version referenced: v2.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096.

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

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481. Association for Computational Linguistics.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024a. o1-Coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154v2*. Version referenced: v2.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024b. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.

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:2411.14405v2*. Version referenced: v2.

Shaolin Zhu, Leiyu Pan, Dong Jian, and Deyi Xiong. 2025. Overcoming language barriers via machine translation with sparse mixture-of-experts fusion of large language models. *Information Processing & Management*, 62(3):104078.

A Professional Translators Introduction

To ensure the quality and linguistic faithfulness of both CoT template construction and human evaluation, we collaborated with the language service division of a large international enterprise and recruited a pool of full-time professional translators.

In total, 30 human translators participated in this project. All translators are native Chinese speakers, hold formal degrees in linguistics, translation, or related fields, and possess certified advanced English proficiency as required by the company's translation workflow. In their routine professional practice, they work with a diverse set of languages, including *English, French, German, Spanish, Arabic, Portuguese, Greece, Malay, Japanese, Korean, Italian, and Russian*. Table 10 summarizes their roles and experience.

Task	# Translators	Experience
Translation	30	3–12 years
Localization	30	3–9 years
Editing	8	5–9 years
Technical Writing	8	5–9 years
Linguistic Testing	10	5–9 years

Table 10: Work position and experience of the 30 human translators involved in this study. # Translators refers to the number of translators with experience in each task, counted from the same group of 30.

From the translator pool, we selected 8 senior professionals (each with over five years of experience) as core experts for CoT template construction. They participated in semi-structured interviews, iterative refinement, and pilot evaluations to consolidate the six CoT reasoning strategies, ensuring that the resulting templates reflect authentic translation workflows while remaining stable and reproducible as structured reasoning instructions.

B Human Evaluation Protocol

Evaluation Criteria. The human evaluation assesses each translation along three dimensions: **Accuracy**, **Fluency**, and **Overall Direct Assessment (DA)**. Accuracy and Fluency are rated on a 5-point Likert scale (1: severe meaning or fluency breakdown; 5: accurate and fluent with no critical errors) indicating the overall quality level in each dimension. In addition, DA is assigned independently on a 0–100 scale following the rubric in Table 11.

Error Taxonomy and Scoring Guideline. To ensure consistent judgments, annotators referred to a predefined error taxonomy (Table 12) adapted from established MQM-style categories (Freitag et al., 2021). The DA score bands and rubric (Table 11) were defined as a task-specific scoring guideline inspired by standard Direct Assessment practice in MT evaluation (Barrault et al., 2019). The taxonomy distinguishes between Accuracy-related (meaning-level) issues and Fluency-related (surface-level) issues.

During evaluation, annotators identified whether a translation issue was present, determined its dimension, and assessed its severity (e.g., minor, major, or critical). These severity judgments guided the assignment of Accuracy and Fluency Likert performance levels and DA score bands. No explicit error counting or weighted MQM aggregation was performed.

DA	Rubric
0–20	Meaning unclear or completely wrong; only small fragments are correct; extremely hard to understand.
21–40	Only very limited meaning preserved; key information missing/wrong; poor readability with many disfluencies/grammar issues.
41–60	Partially conveys meaning but many key errors remain; readability acceptable but lacks naturalness.
61–80	Mostly conveys key meaning; some non-critical errors; may contain grammatical issues or mildly unnatural expressions.
81–100	Meaning largely faithful; only minor non-critical errors; fluent and natural overall.

Table 11: DA scoring rubric used in our human evaluation.

Dimension	Error types
Accuracy	Terminology misuse; lexical mistranslation; named-entity errors; date/time errors; numerical errors; logical errors; over-literal translation; addition; omission; DNT violations; untranslated segments.
Fluency	Spelling/casing/symbol errors; punctuation errors; whitespace/formatting errors; grammar/syntax errors; repetition; locale-specific conventions (numbers, currency, units, dates, names, addresses); style issues (awkward/unnatural phrasing); inconsistency; unintelligible output.

Table 12: Error taxonomy used for human evaluation.