MT-R1-Zero: Advancing LLM-based Machine Translation via R1-Zero-like Reinforcement Learning

Anonymous ACL submission

Abstract

Large-scale reinforcement learning (RL) methods have proven highly effective in enhancing the reasoning abilities of large language models (LLMs), particularly for tasks with verifiable solutions such as mathematics and coding. However, applying this idea to machine translation (MT), where outputs are flexibly formatted and difficult to automatically evaluate with explicit rules, remains underexplored. In this work, we introduce **MT-R1-Zero**, the first open-source adaptation of the R1-Zero RL framework for MT without supervised fine-tuning or cold-start. We propose a rule-metric mixed reward mechanism to guide LLMs towards improved translation quality via emergent reasoning. On the WMT 24 English-Chinese benchmark, our MT-R1-Zero-3B-Mix achieves competitive performance, surpassing TowerInstruct-7B-v0.2 by an average of 1.26 points. Meanwhile, our MT-R1-Zero-7B-Mix attains a high average score of 62.25 across all metrics, placing it on par with advanced proprietary models such as GPT-4o and Claude-3.5-Sonnet, while the MT-R1-Zero-7B-Sem variant achieves state-of-the-art scores on semantic metrics. Moreover, our work exhibits strong generalization capabilities on out-of-distribution MT tasks, robustly supporting multilingual and low-resource settings. Extensive analysis of model behavior across different initializations and reward metrics offers pioneering insight into the critical role of reward design, LLM adaptability, training dynamics, and emergent reasoning patterns within the R1-Zero paradigm for MT. Our code is available at https://anonymous.4open.science/r/MT-R1-Zero-Anonymous.

1 Introduction

Large-scale Reinforcement Learning (RL) has empowered Large Language Models (LLMs) with strong reasoning capabilities (OpenAI, 2024; Team, 2025a,b), demonstrating significant success in tasks such as mathematical reasoning or coding in which answers can be clearly verified. In particular, DeepSeek-R1-Zero (DeepSeek-AI et al., 2025) introduced a pure rule-based RL approach that directly fosters emergent reasoning ability without requirements on structured Chain-of-Thought (CoT) data (Huang et al., 2025; Cui et al., 2025) or sophisticated techniques such as Monte Carlo Tree Search (MCTS) (Silver et al., 2016; Zhao et al., 2024; Luo et al., 2024; Guan et al., 2025). However, the applicability of these methods to machine translation (MT) remains challenging and underexplored, as MT outputs are flexibly generated and hard to evaluate automatically with explicit rules.

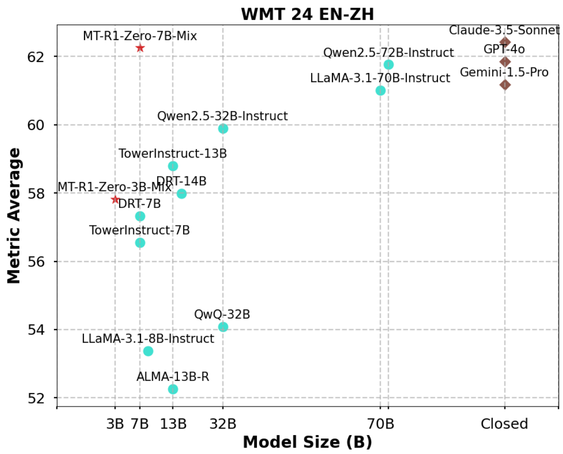

Figure 1: Performance comparison of contemporary LLM-based translation systems on the WMT 24 EN-ZH test set, plotted by average score across BLEU, COMETKiwi, and XCOMET versus model size.

Recent work has launched attempts to empower LLMs for MT with reasoning capabilities (Chen et al., 2025; Liu et al., 2025). Early studies investigate explicit reasoning methods for improved translation, such as finetuning with CoT (Wang et al., 2024a) or MCTS (Zhao et al., 2024) , where advanced multi-step pipelines with self-correction or long-thought agentic mechanisms are further explored (Feng et al., 2024b; Wang et al., 2024b,a).

Another line of work leverages RL to empower LLMs for MT through process reward models or supervised finetuning (SFT) with manually annotated CoT data (Feng et al., 2025; He et al., 2025). However, these methods often depend on manually designed or synthetically generated structured CoT data, rely on complex search algorithms, or require explicit multi-stage prompting, leaving the potential of pure RL-based approaches largely unexplored. Furthermore, the performance reported in these studies often lags behind state-of-the-art (SoTA) open-source or proprietary models.

Developing pure RL methods to directly enhance the reasoning ability of LLMs for better translation requires answering three key questions: 1) **Feasibility**: How to design R1-Zero-like RL pipelines with effective reward signals to directly solve MT tasks without binary rule-based rewards; 2) **Reasoning capability**: Could pure RL training cultivate emergent reasoning abilities and induce models to generate explicit thinking patterns for MT, such as multi-step CoT or verification/reflection; 3) **Generalizability**: Could the training paradigm generalize across different models (e.g., pre-trained base models, instruction-tuned models, or models pre-trained on translation data) or diverse downstream settings (e.g., out-of-distribution, multilingual or low-resource scenarios).

In this work, we introduce **MT-R1-Zero**, the first open-source implementation that extends the R1-Zero-like RL training paradigm to MT. We propose a rule-metric mixed reward mechanism that adapts the original rule-based reward concept to effectively guide training in MT scenarios. We explore different rewards optimizing over lexical (Lex), semantic (Sem), and Lex-Sem mixed (Mix) objectives to guide LLMs towards improved translation quality via emergent reasoning. Our experiments demonstrate the efficacy of this approach: as RL training progresses, our MT-R1-Zero-3B-Mix achieves competitive performance, surpassing TowerInstruct-7B-v0.2 by an average of 1.26 points across all metrics (BLEU, COMETKiwi, XCOMET) on the WMT 24 English-Chinese (EN-ZH) benchmark. Meanwhile, our MT-R1-Zero-7B-Mix surpasses LLaMA-3.1-70B by an average of 1.24 points and Qwen2.5-72B by 0.48 points, even on par with top proprietary models such as GPT-4o and Claude-3.5-Sonnet. The MT-R1-Zero further demonstrates promising generalizability across multilingual and low-resource settings.

Extensive experiments further provide key findings and insight into the adaptation of R1-Zero paradigm to MT. First, we empirically demonstrate that the choice of metric reward plays a pivotal role in steering RL optimization and translation style (semantic or lexical) (Finding 1). Further analysis reveals that MT-R1-Zero induces diverse emergent reasoning patterns, including dynamic language-of-thought transition during translation (Finding 2). We also identify distinct RL adaptability of different LLMs (Finding 3). Ablation studies suggest that pure RL process alone can lead to substantial translation improvements, independent of thinking verbosity. Our core contributions are as follows:

- We present the first open-source implementation of the DeepSeek-R1-Zero paradigm for MT, achieving superior performance across in-domain and out-of-distribution MT tasks.
- Our analysis reveals key findings and recipes for effective R1-Zero adaptation to MT, including reward metric selection, emergent reasoning patterns, training dynamics and LLM adaptability.
- Extensive experiments and ablations show that pure RL serves as the primary driver of MT improvements, with minimal dependence on forced reasoning or output length, highlighting the significant potential of RL for diverse translation applications and broader language tasks.

2 Related Work

LLM Reasoning with Post-training. Recent research indicates that scaling test-time computation can significantly enhance the ability of LLMs to tackle complex reasoning tasks (OpenAI, 2024; Zeng et al., 2024; Xiang et al., 2025). Many approaches rely on sophisticated techniques such as step-level process reward models (PRMs) that provide granular feedback (Lightman et al., 2024; Yuan et al., 2024; Snell et al., 2024) or MCTS to explore potential reasoning paths (Feng et al., 2023; Qi et al., 2024; Guan et al., 2025). A recent alternative, DeepSeek-R1-Zero (DeepSeek-AI et al., 2025), demonstrated that large-scale pure RL, guided only by formatting rules and correctness of final predictions (rule-based reward), can motivate LLMs to develop self-emergent reasoning processes for complex reasoning tasks.This paradigm has been successfully replicated and extended to mathematical, logical, and visual reasoning (Hu et al., 2025; Face, 2025; Xie et al., 2025; Huang et al., 2025). Despite its potential, the application of the R1-Zero RL paradigm to challenging generation tasks like MT, in which the accuracy/quality of outputs is not rule-based and difficult to validate automatically, remains an open question.

LLM Reasoning for MT. Leveraging reasoning to improve MT has garnered increasing attention, as explored by Chen et al. (2025); Liu et al. (2025). Previous work have designed multi-step processes for MT, e.g., Feng et al. (2024b) introduced an API-based self-correcting framework, and Wang et al. (2024b) employed multi-task training followed by a multistage inference phase. Wang et al. (2024a) integrated a similar procedure into inference-time CoT, using a multi-agent mechanism to synthesize long CoT prompts for English-Chinese literary translation. Efforts have also focused on reward modeling for MT reasoning. Feng et al. (2025) constructed implicit process reward models for translation and explored their effectiveness when combined with test-time search. Recent study further evaluated explicit reasoning for MT using CoT fine-tuning and MCTS to expand test-time computation (Zhao et al., 2024). He et al. (2025) demonstrated that models can acquire reasoning-based translation capabilities through multi-stage training with manually constructed CoT templates.

3 Method

In this section, we present our method that trains a translation model with pure RL using a hybrid reward system. Unlike tasks with fixed correct answers, translation allows for multiple valid outputs, making the evaluation more complicated. In this work, we introduce a rule-metric mixed reward that integrates reasoning format checking with multiple translation quality assessment metrics, which is used within the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) algorithm to ensure stable and efficient RL training.

3.1 Rule-Metric Mixed Reward

The reward signal r is crucial in RL. DeepSeek-R1-Zero (DeepSeek-AI et al., 2025) employs simple rule-based rewards that check whether the answer is correct and whether the response follows a specific format. This works well for tasks with fixed format correct answers such as math or coding. However, there is often no single "correct" output for MT, impeding the design of rule-based rewards. Fortunately, the MT community has developed many evaluation metrics to measure translation quality. Recent advancements in automated MT evaluation metrics have shown promise in aligning automated assessments with human translation quality judgments (Freitag et al., 2023). Thus, we design a rule-metric mixed reward, which consists of two parts: a Format Reward that checks output structure, and a Metric Reward that evaluates translation quality. We use a structured prompt template similar to that in DeepSeek-R1-Zero:

Template for MT-R1-Zero

A conversation between User and Assistant. The User asks for a translation from {src_language} to {tgt_language}, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the user with the final translation. The reasoning process and final translation are enclosed within <think> </think> and <translate> </translate> tags, respectively, i.e., <think> reasoning process here </think><translate> final translation here </translate>.
User:{src_text}
Assistant:

Here, `src_language` and `tgt_language` indicate the source and target languages, and `src_text` denotes the source text requiring translation.

Format Reward: We use regular expression extraction to enforce a structured response format. The model is required to place reasoning process within `<think></think>` tags and provide the translation inside `<translate></translate>`. The format reward score (S_{format}) is computed as:

$$S_{format} = \begin{cases} 1, & \text{if format is correct} \\ \text{-1}, & \text{if format is incorrect} \end{cases}$$

Metric Reward: We use automatic evaluation metrics to calculate a translation quality score S_{metric}. We explore three approaches:

1. **N-gram Lexical Matching Reward *(Reward-Lex)*:** Metrics such as BLEU (Papineni et al., 2002) or chrF (Popović, 2015) evaluate translation quality by measuring the difference (primarily lexical overlap) between the translation and the human-written reference. In our experiments, we calculate BLEU via `sacrebleu`[1].

2. **Semantic and Contextual Reward *(Reward-Sem)*:**Learning-based metrics like COMET (Rei

[1]https://github.com/mjpost/sacrebleu

et al., 2020) and COMETKiwi (Rei et al., 2022) are trained on human judgments (e.g., MQM quality assessments (Freitag et al., 2021)). These metrics can recognize good translations even if the wording differs from the reference, as long as the meaning is preserved. We use the COMETKiwi-23-XL, which was used in the WMT 24 (Kocmi et al., 2024) and only needs the source sentence and the model's translation.

3. **Lexical and Semantic Mixed Reward *(Reward-Mix)*:** To capture both lexical fidelity and semantic adequacy, we use a hybrid reward *(Reward-Mix)* that adds together Lexical Matching Reward *(Reward-Lex)* and Semantic and Contextual Reward *(Reward-Sem)*.

Accordingly, the computation of S_{metric} depends on the selected reward configuration:

$$S_{metric} = \begin{cases} \mathrm{B}(\mathrm{trans}, \mathrm{ref}), & \text{if } \textit{Reward-Lex} \\ \mathrm{CK}(\mathrm{src}, \mathrm{trans}), & \text{if } \textit{Reward-Sem} \\ \mathrm{B}(\mathrm{trans}, \mathrm{ref}) + \mathrm{CK}(\mathrm{src}, \mathrm{trans}), & \text{if } \textit{Reward-Mix} \end{cases}$$

where B denotes normalized BLEU score, CK denotes the COMETKiwi score, `trans` is the generated translation, `ref` is the reference translation, and `src` is the source text.

Rule-Metric Mixed Reward: The final reward r combines both the format reward (S_{format}) and the metric reward (S_{metric}). Formally, it is calculated using the following rule:

$$r = \begin{cases} S_{format} - 2, & \text{if } S_{format} = -1 \\ S_{format} + S_{metric}, & \text{if } S_{format} = 1 \end{cases}$$

where S_{metric} is calculated only if the response format is correct $S_{format} = 1$. Then the final reward becomes $r = 1 + S_{metric}$. Unlike traditional rule-based rewards that give a fixed score for correct outputs, our approach uses a continuous metric score. This means the reward can vary within the [1, 2] or [1, 3] range, depending on translation quality. As a result, the model receives more detailed feedback and can learn to improve even small differences in translation quality.

3.2 RL Algorithm

We use the GRPO algorithm (Shao et al., 2024) to train the translation model with our rule-metric mixed reward. In each training step, for a given query q, we sample a group of candidate outputs $\{o_1, o_2, \cdots, o_G\}$ from the policy model $\pi_{\theta_{old}}$. $A_i = \frac{r_i - \mathrm{mean}(\{r_1, r_2, \ldots, r_G\})}{\mathrm{std}(\{r_1, r_2, \ldots, r_G\})}$ is the computed advantage using the group rule-metric mixed rewards $\{r_1, r_2, \cdots, r_G\}$. GRPO then maximizes the following objective function to optimize π_θ:

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

where ε and β are hyperparameters controlling the PPO clipping threshold and the weight of the Kullback–Leibler (KL) divergence penalty (Schulman et al., 2017; Shao et al., 2024), respectively.

4 Experiments

4.1 Experimental Setup

Dataset and Benchmarks. Our experiments primarily focus on English (EN) and Chinese (ZH). Following Xu et al. (2023) and Feng et al. (2024a), we sourced EN⇌ZH parallel examples from WMT 2017-2020, totaling 13,130 pairs. For in-domain evaluation, we use WMT 24 (EN-ZH) and WMT 23 (ZH-EN). Out-of-distribution (OOD) generalization is assessed on benchmarks covering: (1) unseen language pairs from WMT: English-Japanese (EN-JA, WMT 2024) and German-English (DE-EN, WMT 2023 Document-level); and (2) unseen language pair from a distinct dataset: German-Chinese (DE-ZH, Flores-200 (Costa-jussà et al., 2022)). Detailed statistics are in Appendix E.

Baselines. Our primary baselines encompass leading proprietary models, namely Claude-3.5-Sonnet (Anthropic, 2024), GPT-4o (OpenAI, 2023), and Gemini-1.5-Pro (Team et al., 2024), alongside advanced open-source models such as the Qwen2.5 series (Yang et al., 2024), LLaMA-3.1 series (Grattafiori et al., 2024), advanced multilingual models Aya series (Aryabumi et al., 2024), and translation-specific Tower family (Alves et al., 2024). More evaluation details are in Appendix D.

Evaluation Metrics. We assess translation quality using a suite of complementary metrics, including the lexical metric BLEU (Post, 2018), the reference-free learning-based metric COMETKiwi-23-XL (Rei et al., 2022), and a SoTA, reference-based learning metric XCOMET-XL (Guerreiro

Model	ZH-EN				EN-ZH			
	BLEU	COMETKiwi	XCOMET	Avg.	BLEU	COMETKiwi	XCOMET	Avg.
Closed								
Claude-3.5-Sonnet (2024/10)	22.55	71.69	87.32	60.52	38.63	70.39	78.24	62.42
GPT-4o (2024/08)	22.57	71.63	87.22	60.47	41.13	69.01	75.43	61.86
Gemini-1.5-Pro (2025/03)	18.34	69.23	85.55	57.71	39.82	67.47	76.26	61.18
Open								
General Purpose LLMs								
LLaMA-3.1-70B-Instruct	25.19	70.43	86.21	60.61	39.82	68.05	75.17	61.01
Qwen2.5-72B-Instruct	21.96	70.95	87.07	59.99	39.29	69.04	76.97	61.77
Qwen2.5-32B-Instruct	20.54	69.35	85.47	58.45	36.36	68.43	74.90	59.90
Multilingual and Translation-Specific LLMs								
TowerInstruct-13B-v0.1	24.72	70.17	85.69	60.19	37.06	66.22	73.13	58.80
TowerInstruct-7B-v0.2	23.32	69.99	84.93	59.41	34.93	64.04	70.67	56.55
Aya-23-35B	21.99	68.68	84.32	58.33	36.33	64.40	72.10	57.61
Aya-23-8B	19.13	66.74	82.89	56.25	33.28	63.11	70.78	55.72
Ours								
Qwen2.5-3B-Base	14.26	64.86	76.76	51.96	15.90	52.05	67.13	45.03
MT-R1-Zero-3B-Lex	21.53	66.33	81.69	56.52	33.70	60.58	65.67	53.32
MT-R1-Zero-3B-Sem	18.41	70.33	85.98	58.24	24.32	69.75	76.92	57.00
MT-R1-Zero-3B-Mix	22.54	68.84	84.08	58.49	36.27	65.05	72.10	57.81
Qwen2.5-7B-Base	18.23	68.27	84.99	57.16	31.14	63.38	69.83	54.78
MT-R1-Zero-7B-Lex	23.56	65.35	82.12	57.01	40.11	64.57	70.21	58.30
MT-R1-Zero-7B-Sem	16.62	71.66	86.07	58.12	23.07	72.07	79.37	58.17
MT-R1-Zero-7B-Mix	23.98	70.81	86.17	60.32	40.97	69.43	76.36	62.25

Table 1: Performance comparison on in-domain translation directions (EN-ZH, ZH-EN) using BLEU, COMETKiwi, and XCOMET metrics, with average metric scores (Avg.). MT-R1-Zero variants (*-Lex*, *-Sem*, *-Mix*) are compared against closed and open baselines, which are further categorized by accessibility and specialization. The *-Mix* variant often achieves the best balance, while *-Sem* reaches peak semantic scores.

et al., 2024) that was not directly used as an optimization target. To provide an even more holistic view, we also report scores from MetricX-23-XL (Juraska et al., 2023), another high-performing reference-based metric from a distinct framework, and chrF++ (Popović, 2015), which integrates character-level and word-level matching.

Training Details. Our implementation is based on verl[2]. We selected Qwen2.5-base series (3B and 7B parameter variants) as starting models for MT-R1-Zero training. More details are in Appendix F.

4.2 Main Results

In-Domain Performance. As detailed in Table 1 and Table 5, MT-R1-Zero models demonstrate substantial gains over their base versions and achieve competitive performance against existing SoTA benchmarks. For EN-ZH, MT-R1-Zero-7B-Mix surpasses advanced models like GPT-4o and Qwen2.5-72B on average scores. The MT-R1-Zero-7B-Sem variant particularly excels in semantic-level evaluations (COMETKiwi, XCOMET), outperforming strong proprietary models. This strength is further supported by a strong MetricX-23 score of 2.42, notably better than GPT-4o (3.29) and Qwen2.5-72B (3.08). On ZH-EN, MT-R1-Zero-7B-Mix remains highly competitive. MT-R1-Zero-7B-Sem achieves COMETKiwi scores comparable to leading closed models and surpasses strong open-source counterparts. Furthermore, the MT-R1-Zero-3B-Sem delivers impressive performance for its scale. It scores 69.75 COMETKiwi on EN-ZH, which is approximately 1.7 points higher than the much larger LLaMA-3.1-70B and over 0.7 points above Qwen2.5-72B.

Model	Out-of-distribution			
	EN-JA	DE-EN (Doc)	DE-ZH	Avg.
Strong Baseline				
Qwen2.5-72B-Instruct	76.86	89.51	88.42	84.93
LLaMA3.1-70B-Instruct	75.67	88.72	87.42	83.94
Same-size Baseline				
Qwen2.5-7B-Instruct	63.74	87.45	84.43	78.54
LLaMA-3.1-8B-Instruct	64.50	86.84	82.23	77.86
TowerInstruct-7B-v0.2	56.73	89.47	84.28	76.83
Aya-23-8B	72.23	87.97	83.84	81.35
MT-R1-Zero-7B-Lex	60.65	85.25	83.86	76.59
MT-R1-Zero-7B-Sem	71.95	87.68	87.66	82.43
MT-R1-Zero-7B-Mix	68.49	88.69	88.69	81.96

Table 2: Out-of-distribution performance comparison evaluated by the XCOMET metric.

Out-of-Distribution Performance. Our models also demonstrate robust generalization across different OOD scenarios, which encompass unseen WMT language pairs (EN-JA, DE-EN) and the distinct Flores-200 DE-ZH dataset. XCOMET scores presented in Table 2 highlight this: MT-R1-Zero-7B-Sem excels, achieving an average XCOMET score of 82.43 across these tasks, which is 3.89 points higher than Qwen2.5-7B-Instruct. While

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

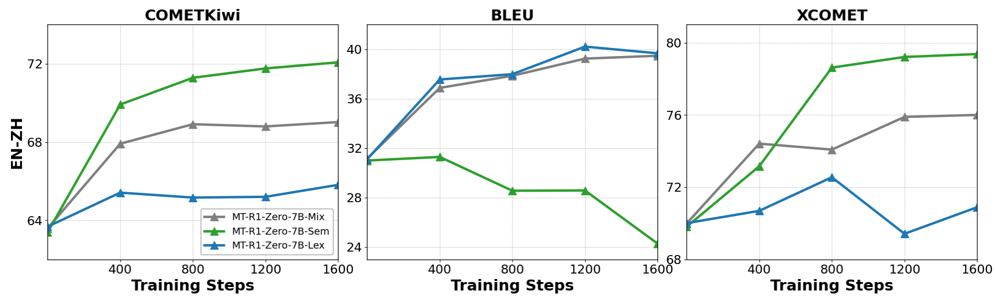

Figure 2: Progression of translation quality metrics using *Reward-Lex*, *Reward-Sem*, and *Reward-Mix*.

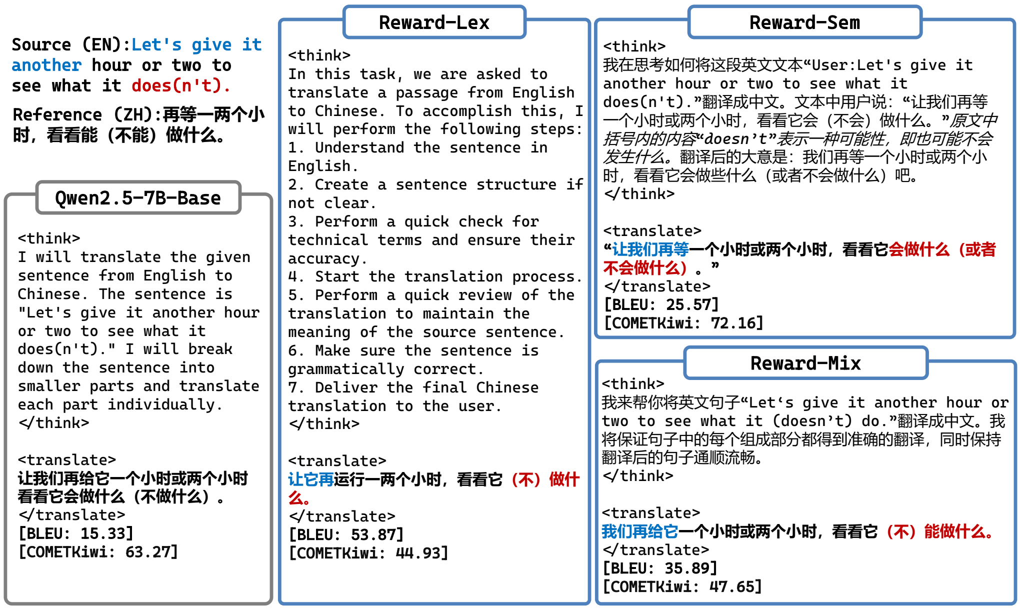

Figure 3: Qualitative examples illustrates the effect of different reward functions (*Reward-Lex*, *Reward-Sem*, *Reward-Mix*) on EN-ZH translation, where the stylistic differences are driven by reward optimization (Finding 1).

our 7B variants do not consistently surpass much larger strong baselines, they significantly outperform other evaluated same-size baselines by a considerable margin, an outperformance that extends to defeating same-size multilingual and translation-specific baselines (such as TowerInstruct-7B-v0.2 and Aya-23-8B). These OOD results suggest that the quality improvements in MT-R1-Zero can effectively transfer to unseen language pairs and benchmarks. More results are provided in Appendix H.

5 Key Findings and Insight

5.1 Impact of Reward Metric Selection

As detailed in Section 3.1, we explore three metric rewards: *Reward-Lex*, *Reward-Sem*, and *Reward-Mix*. Our results demonstrate that the choice among these significantly affects the learning target and final model outputs, as stated in Finding 1.

> **Finding 1:** Reward metric selection critically shapes optimization targets and translation style.

Figure 2 presents the training dynamics with different rewards. Training with *Reward-Lex* maximizes BLEU scores, often at the expense of semantic scores, while *Reward-Sem* maximizes COMETKiwi, leading to a decline in BLEU. Training with *Reward-Mix* improves both metrics. Independent evaluation with XCOMET further supports this finding, showing consistent improvements for Sem and Mix variants while fluctuating for Lex. This finding aligns with the insight from Chen et al. (2025), suggesting that lexical and semantic assessments are complementary, particularly for reasoning-oriented LLMs, and combining them can offer a more comprehensive evaluation signal.

Qualitatively (Figure 3), this optimization alignment manifests as distinct translation styles. *Reward-Lex* encourages literal focused translations, potentially sacrificing nuance. *Reward-Sem* fosters translations that prioritize semantic faithfulness, even if lexically divergent from references. In contrast, the mixed reward yields balanced trans-

Figure 4: Examples illustrating language-of-thought phenomenon, i.e., transition of the internal reasoning language in MT-R1-Zero models. The reasoning language transits from English at Step 0 to target language at Step 1600, indicated by **bold** text across various OOD test pairs (Finding 2).

Figure 5: Comparison of training dynamics for different model families (Qwen2.5, LLaMA-3.1, Tower) undergoing MT-R1-Zero RL training, highlighting differences in adaptability (Finding 3).

lations. This demonstrates that the metric reward fundamentally dictates the nature of the translation quality learned (e.g., semantic v.s. lexical). Therefore, careful metric selection and deliberate fusion are essential for tailoring RL-based MT refinement towards specific and desired translations.

5.2 Emergence and Evolution of Translation Thinking Patterns

As R1-Zero-like training lacks a cold-start phase with predefined reasoning patterns, the observed thinking processes should be emergent and shaped by the RL objective.

> **Finding 2:** Diverse reasoning patterns emerge, varying in style and complexity, and moreover, the internal reasoning language could dynamically transit to target languages even for OOD settings.

Our framework incentivizes a variety of reasoning styles within the <think></think> tags (Figure 11). While some instances include explicit "review/refine" steps, these generally appear as pre-planned components rather than the conversational, iterative self-correction characteristic of the "Aha moment" reported in mathematical reasoning tasks (DeepSeek-AI et al., 2025; Hu et al., 2025). This suggests that while MT-R1-Zero successfully encourages thinking, the complexity and specific nature of emergent reasoning are task-dependent.

Furthermore, we observe a striking and interesting **"language-of-thought" (transition in the language used for internal reasoning**) phenomenon during OOD testing (Figure 4). While base models often use English as default thinking language based on template, MT-R1-Zero models progressively transit to utilize the **target language** of the

Model	In-domain				Out-of-distribution					
	ZH-EN		EN-ZH		EN-JA		DE-ZH		DE-EN (Doc)	
	COMETKiwi	XCOMET	COMETKiwi	XCOMET	COMETKiwi	XCOMET	COMETKiwi	XCOMET	COMETKiwi	XCOMET
Qwen2.5-7B (SFT)	69.29	84.80	67.25	74.29	67.77	65.39	67.01	86.17	67.44	86.74
Qwen2.5-7B (RL w/o thinking)	70.78	86.26	69.62	76.03	68.68	68.77	67.84	86.67	68.31	88.30
Qwen2.5-7B (RL w/ thinking)	70.81	86.17	69.43	76.36	69.27	68.49	70.25	89.25	68.74	88.69

Table 3: Performance comparison of different training paradigms: Supervised Fine-Tuning (SFT) vs. RL with explicit thinking (*RL w/ thinking*) vs. RL without explicit thinking (*RL w/o thinking*). Results shown for in-domain and out-of-distribution tasks support the finding that the RL process itself is the primary driver of gains.

translation task for their reasoning process within the <think></think> block during training (see bold Japanese or Chinese text in step 1600). This dynamic adaptation of the internal "language of thought", conditioned on the task, emerges even without direct supervision on reasoning language.

5.3 Training Dynamics of Different LLMs

The effectiveness and training behavior of MT-R1-Zero are significantly influenced by the base LLM architecture and its initial state (pre-trained vs. instruction-tuned). We compare three distinct model families: general (Qwen2.5 and LLaMA-3.1 series) and translation-specific (Tower family).

> **Finding 3:** LLM architectures exhibit distinct adaptability and effectiveness under MT-R1-Zero, with Qwen showing the highest compatibility in format learning and reasoning generation, while LLaMA and Tower face more challenges and tend towards "format hacking".

As shown in Figure 5, both the translation-specific (Tower) and LLaMA-3.1 models exhibit significantly slower adaptation to the required format compared to Qwen models, as evidenced by their delayed format error reduction. Furthermore, qualitative analysis (Figure 10) reveals that these models often circumvent meaningful reasoning by generating minimal or templated placeholder content in the <think></think> tags, potentially "hacking" the format reward. In contrast, Qwen2.5 models demonstrate stronger adaptability, consistently producing coherent reasoning text within the structured framework. This suggests that architectures like Qwen may possess inherent advantages for integrating structured reasoning via RL, a finding that aligns with prior work on cognitive behaviors (Gandhi et al., 2025). However, even Qwen2.5 models occasionally regress to simplistic one-sentence outputs during reasoning tasks, underscoring the instability of exploration in R1-Zero-like training paradigms.

5.4 Disentangling RL and Explicit Thinking

To determine whether the explicit <think> step or the underlying RL optimization is the primary driver of performance gains, we conducted an ablation study. We compared three paradigms: Supervised Fine-Tuning (SFT), our standard MT-R1-Zero-Mix (RL w/ thinking), and an RL variant without the explicit <think> step (RL w/o thinking).

The results presented in Table 3 reveal a key finding: both RL configurations achieve comparable performance while substantially outperforming the SFT baseline across both in-domain and OOD settings. This demonstrates that the major performance improvements in MT-R1-Zero are primarily driven by the RL framework itself, rather than the mere presence of an explicit reasoning step. This core conclusion is further corroborated by similar findings on the DRT literature translation benchmark (see Appendix I for the full analysis).

6 Conclusion

In this work, we introduced **MT-R1-Zero**, the first successful adaptation of R1-Zero RL framework to MT using a novel rule-metric mixed reward mechanism that combines format enforcement with quality metrics. Our MT-R1-Zero significantly improves translation quality, achieving leading results on multiple benchmarks, i.e., our 3B models compete with much larger open-source models, while our 7B models are on par with advanced proprietary models. The MT-R1-Zero also demonstrates strong OOD generalization and multilingual applicability. Through extensive experiments and analysis, we highlight the significant impact of reward metric choice for optimization, showcase distinct adaptability across different LLMs, and reveal that performance gains are principally from the RL process itself rather than reasoning steps or verbosity, establishing R1-Zero as a viable and potent paradigm for advancing MT. More broadly, our work highlights the great potential of RL for diverse language processing tasks beyond translation.

Limitations

While MT-R1-Zero represents a significant advance, certain limitations remain. The emergent reasoning observed, though diverse, did not achieve the sophisticated iterative self-correction capabilities demonstrated in mathematical reasoning tasks using similar RL or R1-like methods. This discrepancy may reflect fundamental differences in task structure or indicate the need for specialized design in translation tasks. One promising direction would be developing task-specific cold-start datasets for SFT before RL optimization, though this would deviate from the pure RL paradigm we investigated here. Future work could focus on inducing deeper reasoning structures specifically beneficial for the MT task, investigating architectural adapatability across a broader range of LLMs, and developing more appropriate reward mechanisms. Exploring applications to specialized domains (e.g., law and healthcare) and general language processing tasks presents promising opportunities to extend this work.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Anthropic. 2024. Claude 3.5 sonnet.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

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

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

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2023. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*.

Zhaopeng Feng, Ruizhe Chen, Yan Zhang, Zijie Meng, and Zuozhu Liu. 2024a. Ladder: A model-agnostic framework boosting LLM-based machine translation to the next level. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15377–15393, Miami, Florida, USA. Association for Computational Linguistics.

Zhaopeng Feng, Jiahan Ren, Jiayuan Su, Jiamei Zheng, Zhihang Tang, Hongwei Wang, and Zuozhu Liu. 2025. Mt-rewardtree: A comprehensive framework for advancing llm-based machine translation via reward modeling. *arXiv preprint arXiv:2503.12123*.

Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024b. Improving llm-based machine translation with systematic self-correction. *arXiv preprint arXiv:2402.16379*.

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.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.

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

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

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

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

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

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Preliminary wmt24 ranking of general mt systems and llms. *arXiv preprint arXiv:2407.19884*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

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

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

OpenAI. 2023. GPT-4: technical work.

OpenAI. 2024. Introducing openai o1. `https://openai.com/o1/`. Accessed: 2024-10-02.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*.

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.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. 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.

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

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

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, L. Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Kimi Team. 2025a. Kimi k1.5: Scaling reinforcement learning with llms.

Qwen Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning.

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

Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024b. 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.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-though. *arXiv preprint arXiv:2501.04682*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

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

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*.

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

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *arXiv preprint arXiv:2412.14135*.

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. *Preprint*, arXiv:2411.14405.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Progression of Response Length and Performance

By observing the training process, we provide several insights into model adaptation and the emergence of reasoning.

> **Finding 2 (Continued):** Response length initially declines rapidly and then gradually increases as training progresses.

Figure 7 (Right) depicts the pattern in *Finding 2 (Continued)* alongside consistent COMETKiwi improvements (Left). Qualitative analysis (Figure 8) reveals that this length trajectory reflects evolving reasoning strategies. The initial decline corresponds to the model mastering the required format while transitioning from naive decomposition (Step 0) to more efficient, direct translations. The subsequent increase aligns with the development of richer semantic analysis and deeper contextual reasoning within the <think></think> tags (Step 1600).

B KL Penalty Constrains Response Length but Not Quality Gains

We investigate the effectiveness of the KL term in the GRPO objective (Equation 1) on response length and translation quality, as it would regularize the policy by discouraging large deviations from the initial reference model. We conducted experiments without the KL penalty (setting $\beta = 0$, Figure 6), and found that the average response length, after an initial drop, began to fluctuate and trend upward during training. This pattern is consistent with R1-Zero-like results in mathematical tasks (Yu et al., 2025; Yeo et al., 2025). Additional ablation of the KL penalty with COMETKiwi reveals that the improvement of translation quality appears to be largely independent of the thinking verbosity. Significant quality gains were achieved in early-stage training (e.g., before Steps 400) before a substantial increase in response length, even in experiments conducted without the KL penalty. This suggests that performance improvements in the MT-R1-Zero setup could not be attributed solely or primarily to increasing reasoning verbosity.

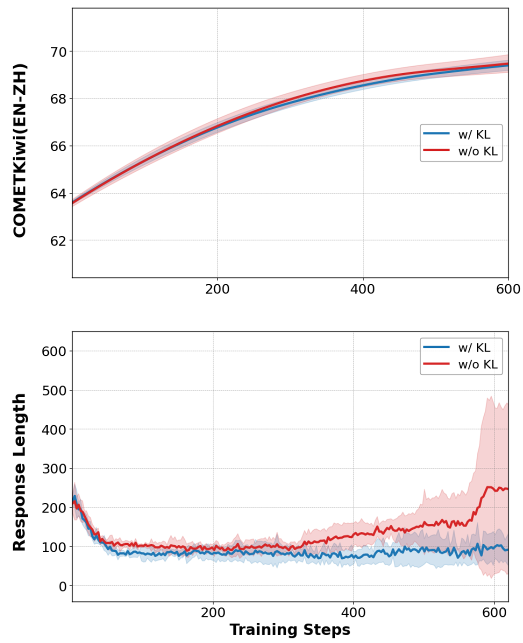

Figure 6: Effect of the KL divergence penalty on EN-ZH COMETKiwi score and response length progression for models trained with (w/ KL, $\beta = 0.01$) and without (w/o KL, $\beta = 0$) the penalty. Experiments are conducted three times with MT-R1-Zero-7B-Sem.

C Multilingual and Low-Resource Support

To evaluate the broader applicability of our framework, we examine its effectiveness in multilingual training scenarios and its potential benefits for low-resource languages. We train multilingual MT-R1-Zero models using the Germanic language data split in the X-ALMA (Xu et al., 2024), augmented with Chinese (see Table 7 for detailed data statistics). We set the batch size to 16 and used COMET-22[3] as the metric reward (Reward-Sem), consistent with the evaluation protocols in X-ALMA. All models are trained for 1 epoch on 16 NVIDIA H800 80G GPUs for about 12 hours. All other hyper-parameters follow the configuration described in

[3] https://huggingface.co/Unbabel/wmt22-comet-da

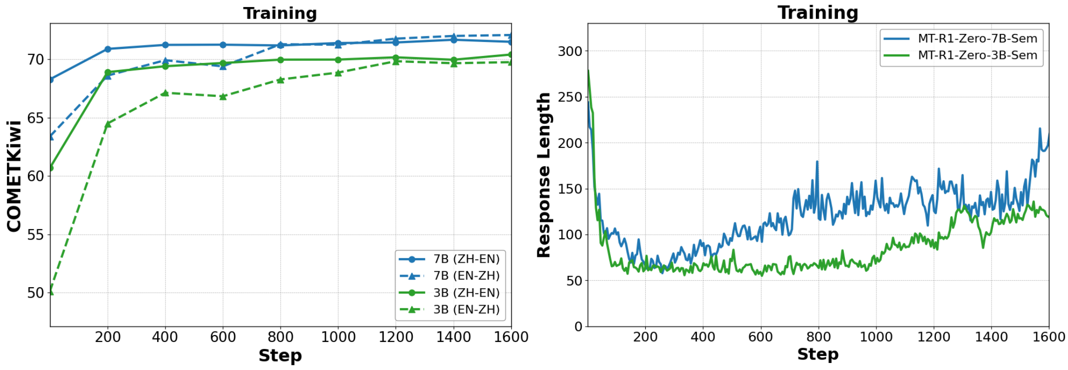

Figure 7: Training dynamics of MT-R1-Zero models (using Reward-Sem). **Left**: COMETKiwi score progression for 3B and 7B models on EN-ZH and ZH-EN test sets. **Right**: Average response length changes over training steps, exhibiting the classic decrease-then-increase pattern (Finding 2 Continued).

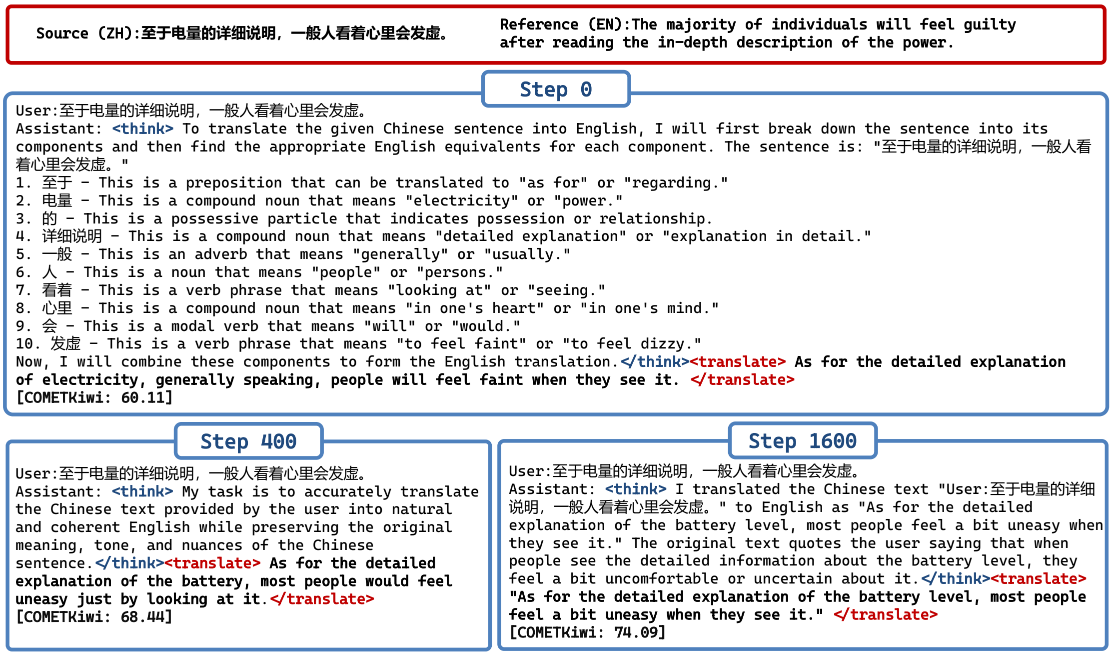

Figure 8: Evolution of an MT-R1-Zero model's reasoning process and translation output for the Chinese source text "其影响可能类似于2008年的经济危机" at different training steps (0, 400, 1600), showcasing the shift from decomposition to more semantic analysis (Finding 2 Continued).

Section 4.1. The training progress, measured by COMET-22 for English-to-target directions, is depicted in Figure 9.

The learning curves demonstrate consistent improvement in translation quality across languages spanning diverse resource levels, including those typically considered low-resource (e.g., Icelandic (IS) and Norwegian (NO)). The steady performance improvement observed throughout training confirms that the MT-R1-Zero framework remains effective when applied in multilingual settings.

D Evaluation Details

When evaluating model performance on the test set, we deployed open-source models locally using frameworks like vLLM[4] or HuggingFace[5] implementations. Proprietary models were accessed via their APIs[6]. We use the sampling decoding strategy with a temperature of 0.2, and top_p set to 0.95. The maximum generation length was capped at 1024 tokens. We adopt the prompt showcasing in Table 4 to sample the translation (applying specific chat template when needed). For the multilingual and translation-specific models, we utilize the prompts in their official model cards.

[4] https://github.com/vllm-project/vllm

[5] https://huggingface.co/docs/transformers/main_classes/text_generation

[6] The specific proprietary models accessed include Anthropic's `claude-3-5-sonnet-20241022`, OpenAI's `gpt-4o-2024-08-06`, and Google's `gemini-1.5-pro`.

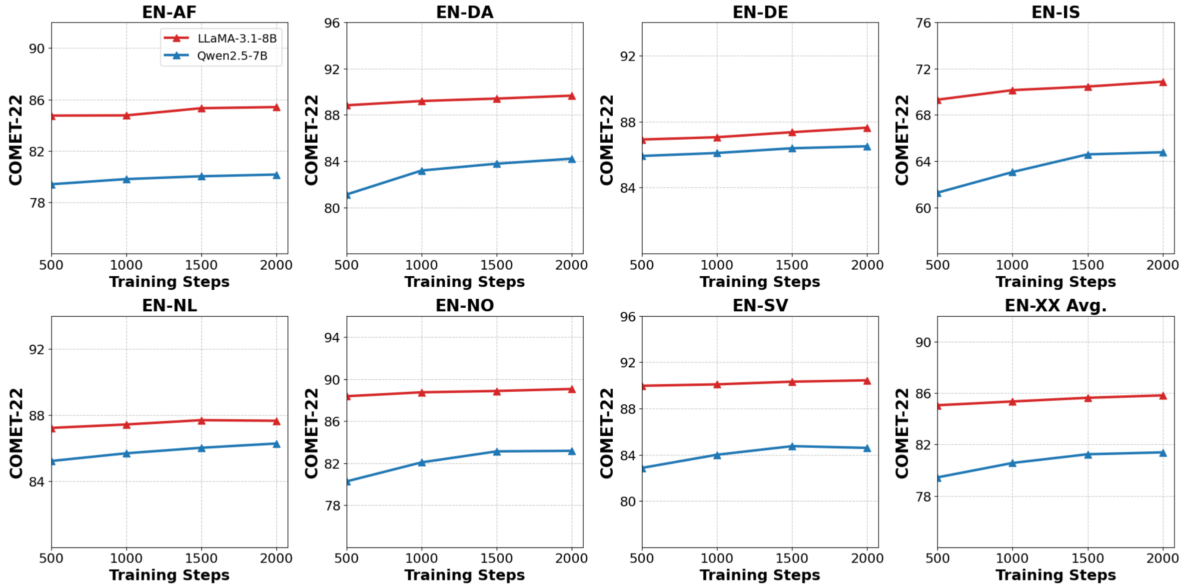

Figure 9: Training progression (COMET-22) for multilingual MT-R1-Zero models based on LLaMA-3.1-8B and Qwen2.5-7B across multiple EN-XX test sets, demonstrating applicability in multilingual settings (Section C).

Inference Prompt
Translate the following text from {src_language} into {tgt_language}. {src_language}:{src_text} {tgt_language}:

Table 4: Prompt used for translation generation. {tgt_language}: target language; {src_language}: source language; {src_text}: the source test sentence.

E Data Statistics

This section provides further details on the datasets used in our experiments. Table 6 outlines the statistics for the data employed in our main EN⇌ZH experiments. The training set for these experiments consists of 13,130 EN⇌ZH parallel sentence pairs, obtained after filtering out sentences with fewer than 30 characters. For model training, these examples were evenly divided between the two translation directions (6,565 pairs each for EN→ZH and ZH→EN). Table 7 presents the statistics for the multilingual experiments discussed in Section C.

F RL Training Details

During training, we configure a batch size of 8 and utilize 8 rollouts per prompt within the GRPO algorithm. We employ a constant learning rate of 5e-7 and set the sampling temperature to 1.0. The maximum generation length for responses is capped at 1024 tokens. We set the KL penalty coefficient β to 0, thereby removing the KL constraint against the reference policy. This decision stems from our empirical observation that the KL penalty tends to restrict the model's exploration of diverse response lengths, which we will discuss further in Section B. The PPO clipping range ϵ is set to 0.2. All models are trained for 1 epoch on 4 NVIDIA H800 80G GPUs for about 13 hours.

Model	ZH–EN		EN–ZH	
	ChrF++ (↑)	MetricX-23 (↓)	ChrF++ (↑)	MetricX-23 (↓)
Closed				
Claude-3.5-Sonnet (2024/10)	50.37	**2.01**	34.51	2.91
GPT-4o (2024/08)	51.06	<u>2.08</u>	32.81	3.29
Gemini-1.5-Pro (2025/03)	45.58	2.21	35.04	3.26
General Purpose LLMs				
LLaMA-3.1-70B-Instruct	51.43	2.32	35.85	3.40
Qwen2.5-72B-Instruct	**52.17**	2.21	**39.27**	3.08
Qwen2.5-32B-Instruct	50.15	2.46	37.47	3.31
Multilingual and Translation-Specific LLMs				
TowerInstruct-13B-v0.1	51.15	2.49	34.17	3.58
TowerInstruct-7B-v0.2	50.17	2.60	32.45	3.96
Aya-23-35B	47.98	2.83	33.33	3.86
Aya-23-8B	44.69	2.86	31.23	3.89
Ours				
MT-R1-Zero-3B-Lex	47.88	3.33	31.74	4.94
MT-R1-Zero-3B-Sem	48.47	2.21	30.27	<u>2.78</u>
MT-R1-Zero-3B-Mix	49.43	2.78	33.12	3.83
MT-R1-Zero-7B-Lex	49.25	3.17	34.27	4.13
MT-R1-Zero-7B-Sem	48.37	2.12	30.57	**2.42**
MT-R1-Zero-7B-Mix	<u>51.55</u>	2.44	<u>38.03</u>	3.11

Table 5: ChrF++ and MetricX-23 scores for all models and translation directions (ZH–EN, EN–ZH). For each column, **bold** marks the best and <u>underline</u> marks the second best performance. (↑): higher is better, (↓): lower is better. MT-R1-Zero rows are aligned with the corresponding variants in Table 1.

G Complementary Metric Scores for Main Results

In addition to the primary metrics discussed in Section 4.2 for in-domain tasks, Table 5 reports complementary scores using MetricX-23-XL and

	Train		Test				
	EN-ZH	ZH-EN	EN-ZH	ZH-EN	EN-JA	DE-EN	DE-ZH
# of cases	6565	6565	997	1976	997	549	1012
Source	WMT 17-20		WMT 24	WMT 23	WMT 24	WMT 23	Flores

Table 6: Data statistics for the training and test sets used in the main experiments (EN⇌ZH).

	Parallel Data				
	Train (from EN)	Train (to EN)	Test (from EN)	Test (to EN)	Resource
Afrikaans (AF)	2994	341	1012	1012	Mid
Danish (DA)	2994	355	1012	1012	Mid
Dutch (NL)	2994	403	1012	1012	High
German (DE)	7015	885	1012	1012	High
Icelandic (IS)	4994	678	1012	1012	Low
Norwegian (NO)	2994	360	1012	1012	Low
Swedish (SV)	2994	339	1012	1012	High
Chinese (ZH)	6906	874	1012	1012	High
English (EN)	-	-	-	-	-

Table 7: Parallel data statistics for languages used in multilingual experiments (Section C), detailing training/test pairs and resource level classification.

chrF++. These metrics provide further perspectives on semantic quality (MetricX-23-XL) and lexical fidelity (chrF++) for the EN⇌ZH evaluations, generally aligning with the strong performance trends of our MT-R1-Zero models.

Model	Out-of-distribution			
	EN-JA	DE-EN (Doc)	DE-ZH	Avg.
Strong Baseline				
Qwen2.5-72B-Instruct	73.25	69.13	69.89	70.76
LLaMA3.1-70B-Instruct	71.84	69.28	68.67	69.93
Same-size Baseline				
Qwen2.5-7B-Instruct	64.79	67.20	67.82	66.60
LLaMA-3.1-8B-Instruct	62.42	66.77	64.28	64.49
TowerInstruct-7B-v0.2	58.33	69.03	65.45	64.27
Aya-23-8B	66.44	67.21	63.40	65.68
MT-R1-Zero-7B-Lex	63.33	66.17	64.32	64.61
MT-R1-Zero-7B-Sem	72.00	68.41	71.51	70.64
MT-R1-Zero-7B-Mix	69.27	68.74	70.25	69.42

Table 8: Out-of-distribution performance comparison using the COMETKiwi metric on EN-JA, DE-EN (Doc), and DE-ZH. (Complements Table 2).

H OOD COMETKiwi and BLEU Scores

To supplement the OOD XCOMET results presented in Section 4.2, we provide COMETKiwi and BLEU scores for the same OOD tasks here. Table 8 details the COMETKiwi results. Table 9 contains the BLEU scores. These additional scores offer further perspectives on the generalization capabilities of our MT-R1-Zero models across the evaluated OOD scenarios.

Model	Out-of-distribution			
	EN-JA	DE-EN (Doc)	DE-ZH	Avg.
Strong Baseline				
Qwen2.5-72B-Instruct	25.02	45.54	40.83	37.13
LLaMA3.1-70B-Instruct	24.64	45.98	37.85	36.16
Same-size Baseline				
Qwen2.5-7B-Instruct	18.91	41.17	35.25	31.78
LLaMA-3.1-8B-Instruct	16.22	40.28	31.08	29.19
TowerInstruct-7B-v0.2	10.52	43.40	34.74	29.55
Aya-23-8B	20.20	37.28	31.86	29.78
MT-R1-Zero-7B-Lex	14.94	40.01	37.00	30.65
MT-R1-Zero-7B-Sem	14.12	33.19	22.83	23.38
MT-R1-Zero-7B-Mix	20.27	43.17	32.31	31.92

Table 9: Out-of-distribution performance comparison using the BLEU metric on EN-JA, DE-EN (Doc), and DE-ZH. (Complements Table 2).

I Details in Disentangling RL and Explicit Thinking

The ablation study discussed in Section 5.4 aimed to determine whether the explicit <think> step or the underlying RL optimization primarily drives performance gains. We compared three training paradigms: 1) **Supervised Fine-Tuning (SFT)**, serving as a non-RL baseline (further SFT training details can be found in Appendix D); 2) our **MT-R1-Zero-Mix (RL w/ thinking)**, which incorporates explicit <think>/<translate> prompting and is trained with the Reward-Mix signal; and 3) an **RL w/o thinking** variant, trained with the same Reward-Mix RL optimization but without prompts enforcing the <think> step.

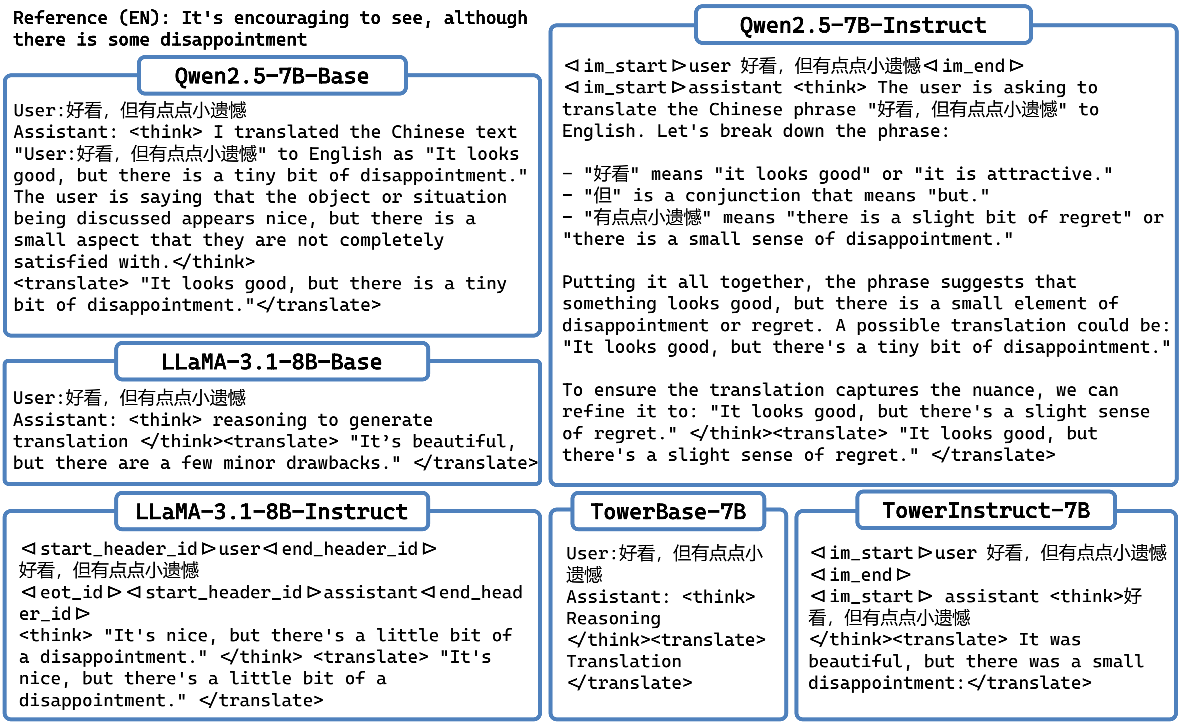

Figure 10: Qualitative comparison of final outputs from different starting models trained with MT-R1-Zero for the Chinese input "好看,但有点点小遗憾", illustrating varying degrees of format adherence and reasoning generation, including format hacking by some models (Finding 3).

Model	DRT Test Set			
	BLEU	COMETKiwi-22	XCOMET	Avg.
Qwen2.5-7B-Instruct	24.17	69.66	61.84	51.89
TowerInstruct-13B	22.71	70.55	62.77	52.01
DRT-7B	35.51	71.77	68.40	58.56
DRT-14B	36.37	72.15	69.64	59.39
Qwen2.5-7B (SFT)	21.61	69.91	63.20	51.57
Qwen2.5-7B (RL w/o thinking)	28.44	72.92	66.17	55.84
Qwen2.5-7B (RL w/ thinking)	28.42	73.20	66.64	56.09

Table 10: Performance comparison on the DRT literature translation dataset (Wang et al., 2024a) using BLEU, COMETKiwi-22, and XCOMET metrics.

The main results from this ablation (Table 3) indicate that both RL configurations ("RL w/ thinking" and "RL w/o thinking") achieve comparable performance and substantially outperform SFT, suggesting the RL framework itself is the main driver of improvements. This core conclusion is further corroborated by evaluations on the DRT literature translation benchmark (Wang et al., 2024a). Table 10 consistently shows that both RL variants perform similarly to each other and are significantly better than SFT on this distinct literary domain. This reinforces our main finding across different benchmarks, underscoring that the RL optimization process, rather than the mere structural prompting for explicit thinking steps, is the key factor behind the enhanced performance of MT-R1-Zero.

J SFT Training Details

For the Supervised Fine-Tuning (SFT) baseline compared in the ablation study (Section 5.4), we utilized LLaMA-Factory (Zheng et al., 2024). The SFT process started from the same base model architecture as the corresponding RL experiments (e.g., Qwen2.5-7B) and was performed on the identical parallel translation dataset (13,130 examples from WMT 2017-2020 after filtering, detailed in Section 4.1). The model was fine-tuned on 8 NVIDIA H800 80G GPUs for 2 epochs using a learning rate of 5e-6 and a batch size of 64, totaling approximately 400 training steps.

I

User:"It's not that bad, right, Kayel?"

Assistant: <think>
When approaching the task of translating the given English text into Japanese, my primary goal is to create a faithful and engaging rendition that maintains the structure, tone, and level of detail of the original while ensuring it reads smoothly and naturally in Japanese. I will pay particular attention to preserving the conversational nature of the exchange and the use of question tags ("right, Kayel?"), as well as the balance between the two characters' statements.

In this case, I will consider the following aspects:

1. **Maintaining the conversational tone**: This involves using appropriate phrasing, word order, and intonation to convey the casual nature of the conversation.
2. **Translating question tags**: Japanese question tags are typically formed differently from English, so I will find an equivalent structure that maintains the flow of the conversation.
3. **Preserving the balance between the two characters' statements**: I will ensure that the translation reflects the equal weight given to both Kayel's response and the speaker's initial statement in the original text.
4. **Adapting to Japanese sentence structure**: Japanese often places the verb at the end of a sentence, so I will adjust the word order accordingly while maintaining the overall meaning and rhythm of the conversation.
</think>
<translate>
"大丈夫じゃないか、ケイエル? "
</translate>

IV

User:好看，但有点点小遗憾
Assistant: <think>
The task at hand involves translating a sentence of Chinese into English while taking into consideration several factors to ensure an effective and appropriate translation. The source text is: “好看，但有点点小遗憾。” Given the context in which this sentence is being used, it is clear that the speaker is expressing their positive opinion of something they have just seen or experienced, but also acknowledging a minor flaw or disappointment that slightly diminishes their overall satisfaction. With this in mind, it is important to maintain the same level of nuance and complexity in the translation to effectively convey the same level of emotional and intellectual engagement expected in professional and polished English communication. After careful consideration of these factors and the specific requirements of the translation context, I have decided to render the source text as follows: "It looks good, but there's a tiny bit of disappointment." This translation not only accurately conveys the same level of appreciation and critique as the source text but also does so in a professional and tactful manner that respects the sensitivity and complexity of the situation at hand.
</think>
<translate>
"It looks good, but there's a tiny bit of disappointment." </translate>

II

User: 北交所融资融券交易的启动是持续推进北交所市场制度建设、完善交易机制、提升二级市场定价功能的重要举措。

Assistant: <think>
When approaching the task of translating the given Chinese text into English, my primary goal is to create a faithful and engaging rendition that maintains the structure, tone, and level of detail of the original Chinese. I will focus on preserving the balance between providing all relevant information and maintaining readability, while also considering the natural flow of the English language. Additionally, I will pay attention to the following aspects:

1. **Structure**: I will aim to maintain the overall structure of the Chinese text, including the use of paragraphs and sentence order, to ensure that the English translation remains coherent and easy to follow.
2. **Tone and voice**: I will strive to capture the tone and voice of the original text, which appears to be informative and objective, while also considering the appropriate level of formality for the target audience.
3. **Terminology**: I will use accurate and up-to-date terminology related to the stock exchange, financing, and trading mechanisms, ensuring that all terms are clear and easily understood by the English-speaking audience.
4. **Cultural and linguistic nuances**: I will pay attention to any cultural or linguistic nuances present in the Chinese text and find suitable ways to convey them in English, where possible, without compromising the clarity or readability of the translation.
5. **Number and date formats**: I will convert any numbers and date formats from the Chinese system to the corresponding English system, as necessary, to ensure that the translation is both accurate and locally appropriate.
6. **Punctuation and formatting**: I will use standard English punctuation and formatting conventions, while taking into account any specific requirements or preferences for the target audience or publication platform.

By considering these aspects, I aim to produce a high-quality English translation that effectively communicates the content and intent of the original Chinese text.
</think>
<translate>
"The launch of financing and margin trading on the Beijing Stock Exchange is an important step in advancing the development of the exchange's market systems, improving trading mechanisms, and enhancing the secondary market's pricing capabilities."
</translate>

V

User: 加拿大也报告了类似的结果，只是数字较低。

Assistant: <think>
When approaching the task of translating the given Chinese text into English, my primary goal is to create a faithful and engaging rendition that maintains the structure, tone, and level of detail of the original Chinese while ensuring it reads smoothly and naturally in English. I will focus on preserving the balance between the two sentences, maintaining the use of parallel structure where appropriate, and accurately conveying the nuances of the Chinese phrasing, including the use of numbers and the comparative nature of the information presented.
</think>
<translate>
"Canada also reported similar results, although the numbers were lower."
</translate>

III

User: to divide the 412.5 meters long, 12.9 meters wide, with a total weight of 15,000 tons of continuous beams into 96 sections.

Assistant: <think> To translate the given English sentence into Chinese, I would follow these steps to ensure accuracy, fidelity to the original text, and naturalness in the target language (Chinese):

1. **Understand the context and meaning of the original text**: I would first read and re-read the sentence to grasp its full meaning, paying close attention to technical terms, measurements, and the overall structure of the information being conveyed. This ensures a proper translation that maintains the original message.
2. **Identify key elements and vocabulary**: I would identify important numbers, measurements, and specific terms that need to be accurately translated, such as "412.5 meters," "12.9 meters," and "15,000 tons." I would also note any technical terms related to engineering or construction, such as "continuous beams" and "sections."
3. **Consider cultural and contextual nuances**: While this particular sentence doesn't contain cultural references that would require adaptation, I would generally be aware of how to translate phrases or expressions that might have different connotations or usages in Chinese culture compared to English.
4. **Choose an appropriate translation strategy**: For this type of technical information, I would likely use a combination of literal translation and necessary modifications to fit Chinese syntax and readability. This might involve reordering words or phrases to sound more natural in Chinese while still maintaining the original meaning.
5. **Translate while considering the audience**: Since this translation is presumably for a Chinese-speaking audience, I would ensure that the translation aligns with how Chinese speakers might express similar ideas, using idiomatic expressions and terminology that would be familiar to them.
6. **Account for register and tone**: Depending on the intended use of the translation (e.g., technical manual, news report, etc.), I would adjust the formality and wording to match the appropriate register. For a technical report, the language would be more formal and precise.
7. **Translate numbers and measurements**: I would carefully translate the numbers and measurements given (412.5 meters, 12.9 meters, 15,000 tons) to their equivalent in Chinese, ensuring that the units of measurement are accurately reflected in the translation (e.g., using "米" for meters and "吨" for tons).
8. **Ensure coherence and flow in the translation**: After translating each part of the sentence, I would review and refine the translation to ensure that it flows naturally in Chinese, making any necessary adjustments to grammar, syntax, and phrasing to create a smooth and coherent translation.
9. **Review and make final adjustments**: Finally, I would review the entire translated sentence for accuracy, fluency, and naturalness in Chinese. I would make any necessary adjustments to improve the translation, checking again against the original text to ensure that all information is preserved and that there are no errors or misunderstandings.

By following these steps, I can provide a translation that is accurate, faithful to the original text, and reads naturally in Chinese.

</think><translate>将总长412.5米、宽12.9米、总重15000吨的连续梁分成96节。</translate>

Figure 11: Qualitative examples (I-V) showcasing the diverse thinking patterns generated by MT-R1-Zero models (Finding 2). In particular, we observe patterns ranging from structured multi-step decomposition (Types I-III) to more colloquial processing (Types IV-V).