

Translation with Thought: Difficulty-Adaptive Reasoning via Reinforcement Learning for Multi-Domain Machine Translation

Yongshi Ye^{1,3}, Biao Fu^{2,3,*}, Chongxuan Huang^{2,3}, Yidong Chen^{2,3}, Xiaodong Shi^{1,2,3,*}

¹Institute of Artificial Intelligence, Xiamen University

²School of Informatics, Xiamen University

³Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism

{yeyongshi, biaofu}@stu.xmu.edu.cn, mandel@xmu.edu.cn

Abstract

Multi-domain machine translation (MDMT) poses a unique challenge due to varying levels of linguistic complexity across domains. Inspired by human translators' ability to adapt reasoning effort based on difficulty, we propose **TwT** (Translation with Thought), a resource-rational framework that learns to modulate inference between intuitive and deliberate reasoning. **TwT** is trained in two stages: (1) supervised fine-tuning on difficulty-aware long chain-of-thought traces distilled from DeepSeek-R1 and rewritten by GPT-4o to reflect human-like reasoning economy, and (2) reinforcement learning with a hybrid reward to optimize translation quality and reasoning efficiency. Evaluated on 15 benchmarks spanning in-domain and out-of-domain settings, as well as 3 seen and 59 unseen languages, with ablations across three backbone models, **TwT-7B** and **TwT-14B** outperform much larger SOTA reasoning models in translation quality, while reducing token usage by 32–60%. These results confirm that aligning translation behavior with cognitive principles enables robust generalization, high translation quality, and efficient reasoning in MDMT.

1 Introduction

Multi-domain machine translation (MDMT) remains a core challenge for language models due to significant variation in terminology, syntax, and style across domains. A key difficulty lies in the uneven distribution of complexity: some inputs are routine, while others require deeper reasoning to resolve ambiguity or domain-specific constructs. However, most MT systems translate all inputs uniformly, lacking mechanisms to adjust inference effort based on domain-specific complexity (Li et al., 2025; Liu et al., 2025). This contrasts with human translators, who adapt reasoning effort to input difficulty (Hvelplund, 2011; Gile and Lei,

2020). They typically rely on fast, intuitive processing (System 1) for familiar content and slower, deliberate reasoning (System 2) when encountering Rich Points (Agar, 1994), such as ambiguous terminology, complex syntax, or cultural disparities. Because the density of such Rich Points varies across domains and correlates with translation difficulty (Lacruz, 2017), current MT systems still largely lack this adaptive reasoning ability.

From the perspective of reasoning allocation, existing MT approaches fall into two extremes. On one end, standard large language model (LLM)-based translators operate purely in System 1 mode: trained via supervised fine-tuning (SFT) on large-scale parallel corpora (Xu et al., 2024a), they produce fluent outputs without explicit reasoning. While efficient, these models struggle with Rich Points and degrade in out-of-domain (OOD) or low-resource settings. Recent efforts have introduced Chain-of-Thought (CoT) prompting into translation (Wang et al., 2025a), but this does not fundamentally solve the problem, because the same reasoning pattern is applied uniformly regardless of input difficulty. Conversely, the emergence of large reasoning models (LRMs), such as DeepSeek-R1 (Guo et al., 2025), represents a shift to the opposite extreme—an overcommitment to System 2. Reinforcement learning (RL) is often used to train these models to generate Long CoT traces, applying reasoning uniformly across inputs. This raises a natural question: *Can RL serve as a bridge to align the model's reasoning trajectory with the human translation process?*

To investigate this, we conduct two preliminary experiments (Section 3). The first examines *Pure RL*, where RL is applied directly to a base model without SFT on annotated reasoning traces. We find that the model rapidly collapses into repetitive, shallow templates, failing to develop domain-specific reasoning behaviors. The second explores *RL with SFT*, which fine-tunes on CoT traces be-

* Corresponding authors.

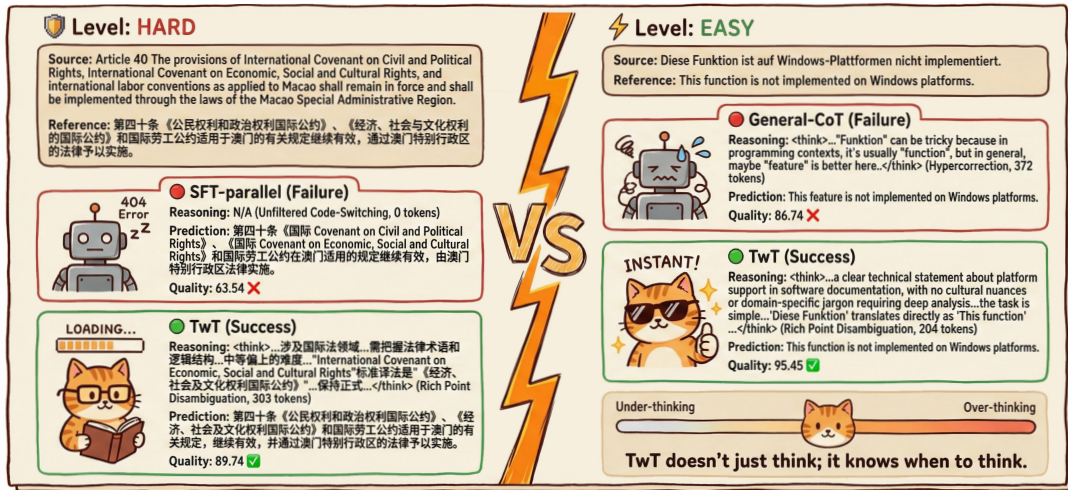


Figure 1: Case Study of Adaptive Thought. TwT switches between System 1 and System 2 based on complexity.

fore RL. While this setup produces longer reasoning, it lacks control over when such reasoning is needed, leading to verbose traces even for simple inputs. This indiscriminate reasoning may help reveal Rich Points, but often results in overthinking and excessive token usage, reducing efficiency and human alignment. Despite this, most reasoning-based MT methods still adopt either *Pure RL* (Feng et al., 2025a) or *RL with SFT* (Wang et al., 2025c), differing mainly in reward design. However, few attempt to align reasoning effort explicitly with input difficulty.

Inspired by human cognitive flexibility, we propose **TwT** (Translation with Thought), a resource-rational framework for MDMT that learns to allocate inference effort based on input difficulty. As shown in Figure 1, TwT dynamically shifts its reasoning behavior according to input difficulty, using concise reasoning for routine inputs and deeper reasoning for domain-specific challenges. To implement this, TwT follows the RL with SFT pipeline. In the cold-start stage, it performs multi-agent distillation to construct difficulty-adaptive reasoning traces across domains: a domain-specialized teacher (DeepSeek-R1) generates diverse reasoning traces, and GPT-4o assesses input difficulty via Rich Points, rewriting the traces to match the appropriate reasoning depth. This process equips our student model with domain-sensitive reasoning and human-like inference modulation, supporting resource-rational translation. In the RL stage, we optimize the adaptive reasoning behavior seeded during cold-start by rewarding high-quality translations. Our hybrid reward combines translation quality metrics (BLEU and COMET) with a

repetition penalty, guiding the model toward efficient, domain-adaptive reasoning through outcome-driven learning.

We conduct a comprehensive evaluation of TwT on 15 benchmarks across in-domain and OOD settings, as well as 3 seen and 59 unseen languages, and perform ablation studies on three different backbone models to assess generalization across both domain and linguistic axes. Our results demonstrate that TwT achieves performance competitive with or superior to SOTA LRMs (e.g., DeepSeek-R1, OpenAI-o1) and surpass strong MT-specialized baselines, while reducing token usage by 32–60%. Empirical analysis confirms that TwT effectively modulates reasoning effort according to task difficulty, leading to more coherent reasoning processes and more accurate translations. This validates the core intuition behind TwT: aligning reasoning effort with input difficulty yields both efficiency and quality gains.

2 Related Work

Recent MT studies increasingly explore explicit reasoning to improve translation quality, starting with shallow strategies such as disambiguation, domain recognition, and self-reflection (Chen et al., 2024; Feng et al., 2025b; Wang et al., 2024b; Hu et al., 2024). To support deeper reasoning, recent studies collect Long CoT traces via MCTS (Zhao et al., 2024) or multi-agent workflows (Wang et al., 2025a), then apply SFT. These traces emulate human translation workflows, improving both performance and interpretability (Chen et al., 2025; Liu et al., 2025). More recently, RL has emerged as a reasoning enhancer (Guo et al., 2025). Sev-

eral approaches optimize translation reasoning with verifiable rewards: R1-T1 uses COMET-based signals (He et al., 2025), MT-R1-Zero combines rule-based and neural metrics (Feng et al., 2025a), DeepTrans employs external LLMs (Wang et al., 2025b), and ExTrans adds exemplar-based guidance (Wang et al., 2025c). However, these methods overlook cognitive alignment; we model human-like reasoning to improve MDMT efficiency and quality.

3 Preliminary Analysis

3.1 Reasoning Collapse

Reasoning Template	Rate
I will translate this Chinese sentence into English by identifying the key phrases and their meanings, and then constructing a coherent English sentence.	42.66%
I will translate this Chinese sentence into English by identifying the key phrases and their corresponding meanings.	24.56%
I will translate this Chinese sentence into English.	5.53%

Table 1: Top-3 reasoning template frequency.

We first investigate the R1-Zero paradigm (Pure RL); detailed training settings are given in Appendix E.1. This setup is motivated by recent findings that RL alone can induce spontaneous reasoning capabilities in math and code tasks (Guo et al., 2025). To test whether this emergence transfers to MDMT, we train models using GRPO with hybrid rewards. To rule out the possibility that KL regularization suppresses exploration (Yeo et al., 2025), we monitor token length dynamics both with and without the KL term. However, unlike the ‘‘Aha moments’’ observed in STEM tasks, our experiments reveal rapid mode collapse.

As shown in Figures 2(a) and 2(b), reasoning traces quickly collapse into shallow patterns (≤ 100 tokens) regardless of the KL setting. Concretely, the model shifts toward high-frequency template recitation, suppressing diverse reasoning. As shown in Table 1, the top three templates account for about 73% of all generated traces in the Zh \rightarrow En direction. This degeneration reveals a key misalignment: MDMT requires domain-aware reasoning, yet without proper initialization, the model produces reasoning that is too brief and overly templated to be elicited reliably. To address this, our cold-start phase explicitly initializes adaptive reasoning behavior before RL.

Model	Easy			Hard		
	Quality	Token	Time	Quality	Token	Time
<i>In-Domain</i>						
SFT-Parallel	64.85	13	12	61.72	40	15
General-CoT	61.54	311	72	56.53	551	80
Domain-CoT	67.36	558	136	62.65	883	149
<i>Out-of-Domain</i>						
SFT-Parallel	67.52	13	7	59.11	36	8
General-CoT	66.59	354	96	60.55	517	97
Domain-CoT	69.81	634	125	62.52	840	138

Table 2: Performance by difficulty for SFT-Parallel, General-CoT, and Domain-CoT. *Quality* is computed as the average of BLEU, COMET, and COMETKIWI. *Time* denotes latency in milliseconds.

3.2 Reasoning Challenges

To evaluate the trade-off between reasoning depth and computational cost, we compare three Qwen2.5-7B-Instruct variants: SFT-Parallel (System 1), General-CoT (System 2), and Domain-CoT, which extends General-CoT with a domain-aware prompt. Details are given in Appendix E.2.

Lack of Domain Awareness. As shown in Table 2, General-CoT performs poorly on in-domain data because its reasoning lacks explicit domain grounding, often defaulting to generic translations rather than domain-specific terminology and fixed expressions. By contrast, SFT-Parallel performs well in these cases by matching the distributional patterns of its training data. This same contrast also explains why General-CoT can be more competitive on OOD data, particularly on harder samples: when the input does not closely match the domain patterns seen in training, intermediate reasoning helps the model better handle syntax, ambiguity, and contextual inference than standard parallel SFT. Importantly, adding explicit domain reasoning largely restores in-domain quality, improving over General-CoT from 61.54 to 67.36 on Easy samples and from 56.53 to 62.65 on Hard samples. This confirms that lack of domain awareness is a major source of in-domain degradation.

Reasoning Redundancy. However, domain awareness alone is not sufficient. General-CoT applies essentially the same reasoning strategy regardless of input difficulty, resulting in substantial redundancy. On Easy samples, this redundancy is clearly detrimental: despite using $23.9\times$ more tokens in-domain (311 vs. 13) and $27.2\times$ more on OOD data (354 vs. 13), it still underperforms SFT-

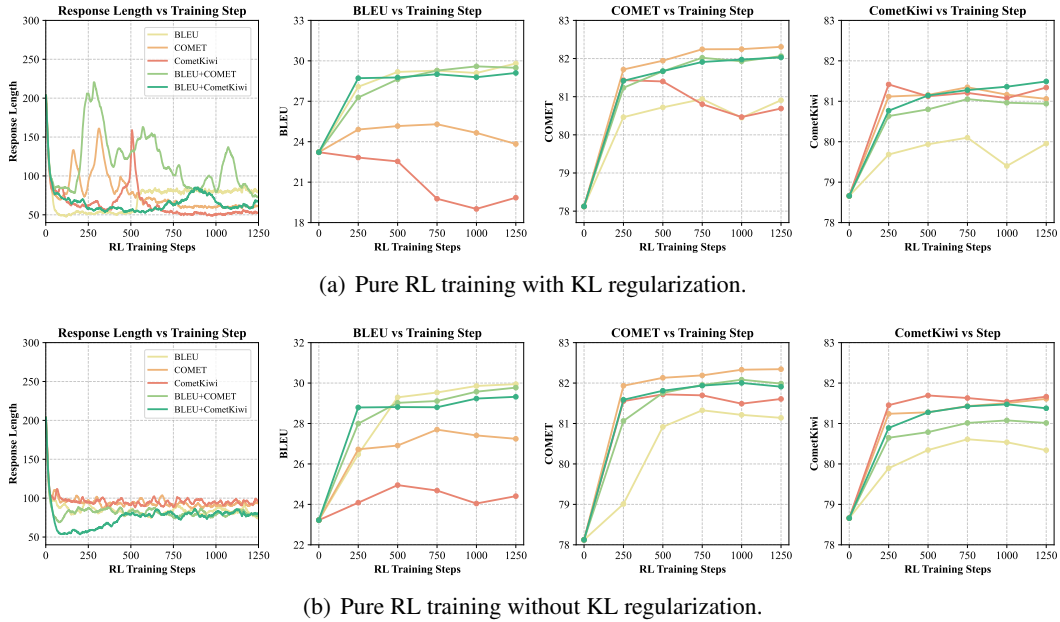


Figure 2: Training dynamics under pure RL using different quality rewards. While translation quality improves under all settings, pure RL training fails to induce extended translation reasoning traces.

Parallel by 3.31 and 0.93 quality points, respectively. Domain-CoT does not solve this problem. Although it restores in-domain quality, it further increases token usage to $42.9\times$ on in-domain Easy samples (558 vs. 13) and $48.8\times$ on OOD Easy samples (634 vs. 13). Similar patterns also hold on Hard samples. These results show that prompt-level domain awareness alone cannot resolve overthinking, motivating **TwT**, which adaptively shifts between System 1 and System 2 to jointly address domain sensitivity and reasoning efficiency.

4 Method

We propose **TwT**, a resource-rational approach that dynamically allocates reasoning effort by input difficulty, mimicking human translation process. As shown in Figure 3, training proceeds in two stages.

4.1 Cold Start

To align the model’s reasoning behavior with the human translation process, we construct a Difficulty-Adaptive CoT Dataset that equips the backbone LLM with adaptive reasoning capabilities. The dataset is curated through a multi-agent distillation-adaptation pipeline. We first prompt DeepSeek-R1 with domain-aware instructions to generate high-quality CoT traces tailored to different domains. These serve as the initial reasoning demonstrations. We then employ GPT-4o, which is verified to align best with human judgment (Ap-

pendix F.7), to assess input difficulty based on the theory of Rich Points. Specifically, difficulty is defined along four linguistic dimensions: sentence complexity, vocabulary rarity, grammatical divergence, and contextual nuance. The corresponding evaluation prompt is shown in Figures 10 and 11.

Conditioned on this assessment, GPT-4o rewrites raw traces into adaptive formats: Easy inputs are reformulated into concise System 1 checks to minimize token usage, whereas Hard inputs retain comprehensive System 2 deliberations for structural and terminological verification. We then perform SFT on the backbone LLM using this compact dataset ($\sim 7k$ examples). We define reasoning depth as the length of the generated reasoning trace (i.e., number of tokens). This process establishes an initial policy distribution over difficulty-aware reasoning strategies, enabling the model to autonomously modulate its reasoning depth. Representative examples are shown in Figures 13 and 14. The resulting data efficiency makes our approach particularly suitable for low-resource settings, demonstrating that robust adaptive reasoning can be achieved with a modest CoT-SFT seed dataset.

4.2 RL Training

To scale this adaptive reasoning behavior, we adopt the GRPO algorithm with hybrid quality rewards, which serve as outcome-driven constraints that encourage the model to identify Rich Points and al-

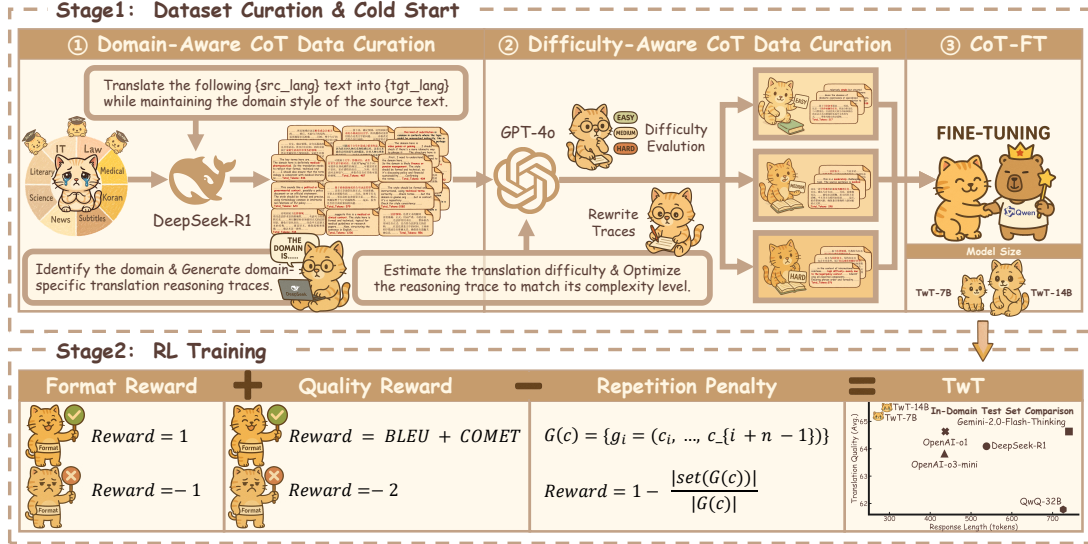


Figure 3: Overview of **TwT** training pipeline. **TwT** is first fine-tuned on difficulty-adaptive Long CoT traces distilled from DeepSeek-R1 and rewritten by GPT-4o for cognitive alignment. RL is then applied with a hybrid reward.

locate deep reasoning selectively—only where it leads to measurable quality improvements. By mitigating indiscriminate overthinking (Section 3.2), we improve reasoning efficiency and align the model with the resource-rational principle introduced in Section 1. The final reward r is crafted from three components to ensure alignment:

$$r = r_f + r_q - \lambda \cdot r_{\text{rep}}$$

Format Reward (r_f). We employ a binary format reward ($r_f \in \{1, -1\}$) to strictly enforce the reasoning-translation structure defined in Figure 5.

Hybrid Quality Reward (r_q). To prevent the model from producing plausible but functionally ineffective reasoning, we introduce a verifiable hybrid feedback signal. Building upon our preliminary analysis in Section 3.1, we identified a critical reward hacking phenomenon. As illustrated by the red curve in Figure 2(a), optimizing solely for a semantic metric (CometKiwi/COMET) leads to a significant metric divergence: despite high semantic reward scores, the lexical accuracy (BLEU) degrades rapidly during training. This confirms that the model hacks the reward by generating vague paraphrases or copying source tokens to maximize semantic similarity, effectively abandoning translation fidelity. To remedy this and strictly enforce alignment, we adopt a hybrid quality reward that combines BLEU (B) with COMET (C):

$$r_q = \begin{cases} B(\hat{y}, y) + C(x, \hat{y}, y) & \text{if } r_f = 1 \\ -2 & \text{if } r_f = -1 \end{cases}$$

This design effectively stabilizes the optimization process. As evidenced by the training dynamics of our **TwT** models (Figure 8(a)), both the 7B and 14B models (Figures 8(b) and 8(c)) exhibit synchronous improvements in lexical and semantic metrics without divergence, validating that the hybrid signal successfully grounds the reasoning process in accurate translation outcomes.

N-gram Repetition Penalty (r_{rep}). We penalize redundant reasoning to discourage degenerate loops and promote efficient token usage. For a given reasoning trace c , let $G(c)$ denote the list of contiguous n -grams. In our experiments, we set $n = 20$. We compute the ratio of repeated n -grams to discourage repetitive, loop-like patterns:

$$r_{\text{rep}} = 1 - \frac{|\text{set}(G(c))|}{|G(c)|} \in [0, 1]$$

5 Experiments

5.1 Experimental Settings

Dataset. We use two datasets for training: (1) a curated set of 7K difficulty-adaptive Long CoT examples spanning 10 domains and three translation directions (De→En, En→Zh, Zh→En) for cold-start SFT, and (2) a separate 20K-sample dataset for RL, constructed from multi-domain parallel corpora. For evaluation, we adopt both in-domain test sets and diverse OOD benchmarks, and additionally include multilingual test sets covering both seen and unseen language pairs. Full dataset details are provided in Appendix B.

Method	Laws		News		Science		Subtitles		Literary		IT		Koran		Medical		Average	
	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens
<i>Large Language Models</i>																		
DeepSeek-V3	77.72	-	69.41	-	69.09	-	62.93	-	56.83	-	66.97	-	57.73	-	69.16	-	66.23	-
Gemini-2.0-Flash	76.57	-	<u>69.30</u>	-	68.80	-	62.96	-	57.41	-	66.54	-	58.13	-	70.21	-	66.24	-
GPT-4o	73.77	-	68.61	-	68.09	-	62.92	-	57.47	-	66.37	-	57.78	-	69.34	-	65.54	-
<i>Large Reasoning Models</i>																		
DeepSeek-R1	77.72	577	68.47	498	68.46	478	61.81	514	53.95	574	66.28	593	57.48	790	69.01	667	65.40	586
Gemini-2.0-Flash-Thinking	76.21	702	68.24	1149	68.17	1092	61.98	708	57.23	781	66.20	345	58.13	677	69.73	415	65.74	734
OpenAI-o3-mini	71.61	428	68.18	443	68.01	385	62.42	355	56.97	546	65.99	343	56.42	511	68.40	346	64.75	420
OpenAI-o1	73.85	478	68.67	408	68.60	367	62.62	340	57.28	521	66.45	403	57.94	506	69.20	441	65.58	433
GPT-5	76.28	784	69.07	740	68.58	606	62.74	519	56.27	859	66.57	492	58.71	751	70.14	531	66.05	660
QwQ-32B	71.84	667	67.88	584	67.86	563	61.80	584	55.09	863	61.44	583	55.81	963	66.68	735	63.55	693
<i>MT-Specialized Models</i>																		
SFT-Parallel-7B	76.58	-	66.08	-	66.38	-	62.87	-	55.96	-	68.02	-	58.05	-	70.21	-	65.52	-
ALMA-7B-R	67.88	-	63.37	-	62.77	-	59.38	-	54.52	-	64.20	-	55.04	-	67.46	-	61.83	-
ALMA-13B-R	70.11	-	64.65	-	64.23	-	60.22	-	55.30	-	64.54	-	55.71	-	68.40	-	62.89	-
TowerInstruct-7B-v0.2	73.91	-	65.93	-	65.45	-	61.07	-	55.32	-	66.78	-	50.05	-	<u>70.73</u>	-	63.66	-
TowerInstruct-13B-v0.1	74.65	-	66.90	-	66.20	-	61.89	-	56.09	-	67.18	-	49.93	-	71.49	-	64.29	-
CoT-FT-7B	76.72	51	66.24	42	66.07	39	62.76	29	55.57	52	67.82	35	57.57	45	70.38	46	65.39	42
MT-R1-Zero-7B	68.93	72	67.41	64	66.98	61	61.97	55	55.80	69	65.60	56	55.52	71	64.43	71	63.33	65
SSR-X-Zero-7B	69.59	56	66.00	52	66.77	49	61.89	39	55.54	54	61.22	36	55.53	46	63.99	50	62.56	48
mExTrans-7B	70.11	597	65.49	553	66.13	546	59.52	476	54.40	610	60.50	452	55.41	604	63.00	565	61.82	551
<i>Our Models</i>																		
TwT-Qwen2.5-7B-Instruct	75.35	310	68.42	311	68.24	294	63.03	247	57.86	281	68.07	222	58.92	269	70.29	262	<u>66.27</u>	274
TwT-Qwen2.5-14B-Instruct	76.58	320	68.62	285	68.32	272	<u>62.97</u>	241	58.14	354	68.37	234	59.35	336	70.59	287	66.62	291

Table 3: In-domain translation performance across eight domains, averaged over En→Zh, Zh→En, and De→En. The **Bold** and underlined values denote the highest and second highest scores, respectively.

Method	Conversation		Ecommerce		Social		Culture		CommonSense		Average	
	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens
<i>Large Language Models</i>												
DeepSeek-V3	68.42	-	<u>66.34</u>	-	66.10	-	69.65	-	65.96	-	67.29	-
Gemini-2.0-Flash	68.77	-	66.26	-	66.10	-	<u>69.02</u>	-	65.23	-	67.08	-
GPT-4o	<u>68.75</u>	-	66.50	-	66.10	-	69.01	-	<u>65.89</u>	-	<u>67.25</u>	-
<i>Large Reasoning Models</i>												
DeepSeek-R1	67.06	534	64.45	552	64.11	554	68.16	560	64.34	602	65.62	561
Gemini-2.0-Flash-Thinking	68.36	1204	65.83	822	65.50	1081	68.42	1220	65.44	2335	66.71	1332
OpenAI-o3-mini	68.04	290	65.87	363	<u>65.58</u>	372	67.55	596	64.45	436	66.30	411
OpenAI-o1	68.47	327	65.80	399	65.38	405	67.96	542	64.78	392	66.48	413
GPT-5	68.40	448	65.49	609	65.14	652	68.33	984	64.09	530	66.29	645
<i>MT-Specialized Models</i>												
SFT-Parallel-7B	65.64	-	63.33	-	62.54	-	65.34	-	61.03	-	63.58	-
ALMA-7B-R	64.71	-	62.54	-	62.68	-	66.63	-	62.02	-	63.71	-
ALMA-13B-R	66.03	-	63.37	-	63.51	-	60.94	-	62.91	-	63.35	-
CoT-FT-7B	65.43	31	63.26	45	62.06	42	64.66	54	61.08	33	63.30	41
MT-R1-Zero-7B	66.59	53	64.16	66	63.69	65	66.23	79	62.32	51	64.60	63
SSR-X-Zero-7B	65.70	37	63.46	50	63.31	49	64.35	66	62.18	34	63.80	47
mExTrans-7B	63.47	464	61.74	566	61.20	555	65.11	631	59.92	470	62.29	537
<i>Our Models</i>												
TwT-Qwen2.5-7B-Instruct	67.72	231	65.71	273	65.48	269	67.82	352	64.48	219	66.25	269
TwT-Qwen2.5-14B-Instruct	67.77	240	65.79	309	65.47	298	68.55	333	64.72	259	66.46	288

Table 4: OOD translation performance across five domains, averaged over En→Zh, Zh→En, and De→En.

Implementation Details. For cold start, we use LLaMA-Factory¹ (Zheng et al., 2024) with Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, and Gemma-2-9B-IT as backbones. We train them on the 7K difficulty-adaptive Long CoT examples for 1 epoch with full-parameter optimization on 8 NVIDIA A100 80GB GPUs, using AdamW with a learning rate of $1e-5$, a total batch size of 32, a cosine learning rate scheduler, a warm-up ratio of 0.1, a maximum input sequence length of 4096, and DeepSpeed ZeRO Stage 3. The cold-start stage completes within 10 minutes. For RL, we use ver1² (Sheng et al., 2025) and train for 1 epoch on 8 NVIDIA A100 80GB GPUs with a total batch size of 16, rollout number 16, rollout temperature 1.0, learning rate $1e-6$, KL loss coefficient $\beta = 1e-3$, maximum response length 2048, and repetition penalty with $n = 20$. RL training

takes about 10 hours. During inference, we use vLLM³ (Kwon et al., 2023) for efficient decoding with temperature 0.0 and repetition penalty 1.05.

Metrics. We report *Quality*, defined as the average of BLEU, COMET (Rei et al., 2020), and CometKiwi (Rei et al., 2022), and *Tokens*, the average length of the generated CoT. The full metric breakdowns are provided in Appendix I.

Baselines. We compare our method against three categories of models: (1) *general-purpose LLMs* such as DeepSeek-V3 (DeepSeek-AI et al., 2024), Gemini-2.0-Flash (DeepMind, 2024), GPT-4o (OpenAI et al., 2024), and open-source models like LLaMA3.1-8B-Instruct (Grattafiori et al., 2024), Gemma-2-9B-IT (Gemma Team et al., 2024), and Qwen2.5 series (7B, 14B, 32B) (Yang et al., 2024). (2) *reasoning-oriented LRMs*, such as DeepSeek-R1 (Guo et al., 2025), Gemini-

¹<https://github.com/hiyouga/LLaMA-Factory>

²<https://github.com/volcengine/ver1>

³<https://github.com/vllm-project/vllm>

Method	En→Zh		Zh→En		De→En		En→X		X→En		Average	
	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens
<i>Large Language Models</i>												
Qwen2.5-7B-Instruct	67.87	-	50.91	-	62.90	-	37.61	-	56.99	-	55.25	-
Gemma-2-9B-IT	66.50	-	52.64	-	60.64	-	54.28	-	66.38	-	60.09	-
<i>MT-Specialized Models</i>												
ALMA-7B-R	64.08	-	54.52	-	62.89	-	44.68	-	44.71	-	54.18	-
Tower-Plus-9B	69.85	-	57.41	-	67.04	-	46.46	-	63.02	-	60.76	-
SFT-Parallel-7B	69.44	-	55.96	-	66.15	-	32.44	-	55.70	-	55.94	-
MT-R1-Zero-7B	67.52	62	55.80	69	62.56	66	40.98	358	58.33	74	57.04	126
SSR-X-Zero-7B	68.55	50	56.52	54	62.86	45	40.43	306	58.19	48	57.31	101
mExTrans-7B	66.70	537	54.40	610	60.74	544	42.72	1047	55.99	731	56.11	694
<i>Our Models</i>												
TwT-Qwen2.5-7B-Instruct	69.99	298	57.86	281	66.45	256	41.13	483	58.84	328	58.85	329
TwT-Gemma-2-9B-IT	69.07	227	58.12	249	66.95	218	54.65	280	67.04	257	63.17	246

Table 5: Results on seen and unseen language directions. En, Zh, and De are *seen* languages, while X denotes *unseen* languages; En→X and X→En report averages over English↔unseen-language directions.

Method	In-Domain		Out-of-Domain	
	Quality	Tokens	Quality	Tokens
TwT-Qwen2.5-7B-Instruct	64.77	278	66.28	263
w/o RP	64.52	301	66.12	271
w/o RP + w/o Adaptive CoT	64.68	748	66.17	720
w/o RP + w/o Cold Start	64.18	62	65.29	59
SFT only w/ Adaptive CoT	62.47	253	64.09	219
SFT only w/ domain-aware CoT	61.65	479	63.47	452
SFT only w/ general CoT	61.39	531	63.11	519
<i>Backbone Models</i>				
Qwen2.5-7B-Instruct	61.99	-	64.71	-
Llama-3.1-8B-Instruct	59.84	-	63.33	-
Gemma-2-9B-IT	59.93	-	63.68	-
<i>Our Models</i>				
TwT-Qwen2.5-7B-Instruct	64.77	278	66.28	263
TwT-Llama-3.1-8B-Instruct	63.20	305	64.44	298
TwT-Gemma-2-9B-IT	64.71	231	66.05	220

Table 6: Ablation study on in-domain and OOD translation test sets. Results are averaged at the dataset level for each setting. RP = repetition penalty.

2.0-Flash-Thinking (DeepMind, 2025), OpenAI o1 (Jaech et al., 2024), o3-mini (OpenAI, 2025b), GPT-5 (OpenAI, 2025a), and QwQ-32B (Qwen-Team, 2025); (3) *MT-specialized models*, including non-reasoning LLMs such as TowerInstruct (Alves et al., 2024), ALMA-R (Xu et al., 2024a,b), SFT-Parallel (Qwen2.5-7B-Instruct fine-tuned on 27K parallel pairs), CoT-FT (Hu et al., 2024), and Tower-Plus-9B (Rei et al., 2025); with reasoning-oriented models including MT-R1-Zero (Feng et al., 2025a), mExTrans (Wang et al., 2025c), and SSR-X-Zero (Yang et al., 2025). For fairer comparison, Appendix C details the training data size of MT-specialized baselines.

5.2 Cross-Domain Generalization

In-Domain. As shown in Table 3, **TwT-14B** achieves a SOTA average quality score of 66.62, outperforming both strong open-source LRMs such

as DeepSeek-R1 (65.40) and closed-source models including GPT-5 (66.05). Notably, our smaller variant, **TwT-7B**, also surpasses dedicated MT systems, demonstrating that our method scales effectively with model size. Compared to mExTrans-7B, which is also trained under the R1 paradigm, **TwT-7B** yields a substantial improvement of +4.45 points while reducing reasoning overhead by 50.27%, highlighting the superior efficiency of our difficulty-adaptive mechanism. The benefits of adaptive reasoning are especially evident in structurally complex and culturally nuanced domains. In the Literary domain, **TwT-14B** consistently outperforms three representative paradigms: it exceeds the System 1 baseline SFT-Parallel by +2.18 points, the heavy-reasoning System 2 model DeepSeek-R1 by +4.19 points, and the pure RL-based MT-R1-Zero-7B by +2.34 points. These results suggest that, for multi-domain translation, neither shallow System 1 execution, indiscriminate System 2 overthinking, nor unguided pure RL exploration alone yields optimal performance. In contrast, our model autonomously modulates reasoning depth to strike a more effective balance between literal accuracy and stylistic adequacy.

OOD. **TwT-7B** achieves a strong average score of 66.25 on five OOD test sets, surpassing MT-specialized baselines such as SFT-Parallel-7B (63.58), which exhibit poor generalization under domain shift. Moreover, **TwT-7B** exhibits a favorable quality-efficiency trade-off, surpassing DeepSeek-R1 while reducing reasoning overhead by 292 tokens. Even in unfamiliar domains, it avoids excessive deliberation by leveraging compact, internalized translation procedures. These results indicate that **TwT** captures domain-agnostic

translation logic rather than relying on domain-specific memorization, approaching the SOTA performance of GPT-4o at a fraction of the computational cost and model size.

5.3 Multilingual Generalization

On seen directions, **TwT-7B** improves Zh→En performance by +6.95 over its base model and +1.90 over SFT-Parallel-7B, under the same training data. On unseen directions (En→X)⁴, **TwT-Gemma-2-9B-IT**, trained on only 27K examples, outperforms the multilingual Tower-Plus-9B, built on the same backbone but trained on 286K examples, by a substantial margin of +8.19. Overall, **TwT** achieves the highest average score across all directions, indicating that its reasoning mechanism generalizes beyond language boundaries and captures transferable alignment strategies.

5.4 Ablation Study

We conduct a comprehensive ablation study to assess the contribution of each component in the **TwT** framework and validate its generalizability across backbone architectures (Table 6). Removing the repetition penalty (w/o RP) leads to a slight quality drop and longer outputs, indicating its role as a regularizer rather than a performance driver. In contrast, removing both the repetition penalty and the difficulty-adaptive rewriting (w/o RP + w/o Adaptive CoT) results in comparable quality but significantly increases reasoning length (from 278 to 748 tokens), highlighting the critical role of adaptive rewriting in controlling verbosity and ensuring inference efficiency. We further examine the necessity of the two-stage training pipeline. Eliminating the cold-start SFT phase (w/o RP + w/o Cold Start) causes performance degradation and length collapse (to 62 tokens), suggesting that RL alone fails to induce structured reasoning behavior. To isolate the impact of SFT data quality, we compare three variants: difficulty-adaptive CoT yields the best result (62.47), followed by domain-aware (61.65) and general CoT (61.39), showing a clear performance hierarchy. Still, only the full **TwT** pipeline achieves the highest score (64.77), confirming that RL is indispensable for turning the adaptive reasoning patterns from mere imitation into an internalized and optimized translation strategy. Lastly, we apply **TwT** to three backbone models: Qwen2.5-7B,

⁴En→X and X→En are averaged over 59 unseen languages from FLORES+; see Appendix B.6 for the full list.

Llama-3.1-8B, and Gemma-2-9B, and observe consistent improvements. For instance, **TwT** improves Gemma-2-9B-IT from 59.93 to 64.71 in-domain and from 63.68 to 66.05 OOD, demonstrating that **TwT** is a model-agnostic framework that robustly enhances translation reasoning regardless of the underlying architecture.

6 Empirical Analysis

6.1 Human Reasoning Alignment

To benchmark **TwT**'s reasoning against human cognition, we conduct a qualitative analysis on 10 Zh→En examples, with expert commentary from a translation studies faculty member. A representative case is shown in Appendix G.2.

Cognitive Convergence. The analysis revealed that **TwT**'s reasoning exhibits strong parallels with human translators in early-stage decision-making. Specifically, **TwT** effectively (1) identifies translation domain and stylistic register, (2) handles complex sentence structures with appropriate syntactic parsing, and (3) demonstrates context-aware terminology adaptation. For example, it consistently distinguishes between literary and technical expressions and adjusts lexical choices accordingly. Its structured CoT mirrors key aspects of professional reasoning—such as coherence maintenance, discourse flow control, and sensitivity to stylistic norms—indicating that **TwT** has internalized domain-aware reasoning behavior resembling human translation logic.

Pragmatic Divergence. Despite these strengths, **TwT** still shows gaps compared with expert translators. It occasionally struggles with cross-sentence consistency in terminology, especially when handling long-form repetitions or abbreviated references. Moreover, its output lacks fine-grained control over tone, idiomaticity, and cultural adaptation, which human translators adjust based on pragmatic context and target audience. These issues suggest that **TwT**'s reasoning remains less flexible in discourse-level adaptation, reflecting the absence of high-level pragmatic awareness. Future work seeks to bridge these gaps by integrating process-oriented feedback, thereby fostering deeper pragmatic alignment with human cognitive processes.

6.2 Reasoning Redundancy Reduction

To assess whether **TwT** eliminates unnecessary computation, we employ DeepSeek-V3.2 to detect six

Redundancy Type	Resolved (%)	Not Resolved (%)
Over-segmentation	87.3	12.7
Unnecessary linguistic explanation	93.2	6.8
Semantic repetition	96.9	3.1
Irrelevant information	93.7	6.3
Redundant alternative translations	96.5	3.5
Low-density long descriptions	97.2	2.8

Table 7: Resolution rates of six redundancies by TwT.

Method	Easy		Medium		Hard		All	
	Quality	Tokens	Quality	Tokens	Quality	Tokens	Quality	Tokens
DeepSeek-R1	66.30	480	63.82	556	62.15	579	64.09	538
Gemini-2.0-Flash-Thinking	66.62	432	64.34	754	62.93	1035	64.63	740
OpenAI-o3-mini	66.57	280	63.44	430	61.44	596	63.82	435
OpenAI-o1	67.14	302	64.26	432	62.49	578	64.63	437
QwQ-32B	62.91	577	61.92	758	60.52	844	61.78	726
General-CoT	61.54	311	60.27	433	56.53	551	59.45	432
TwT-Qwen2.5-7B-Instruct	67.93	216	64.85	272	62.78	332	65.19	273
TwT-Qwen2.5-14B-Instruct	67.98	207	65.23	300	63.26	378	65.49	295

Table 8: In-domain performance by difficulty level.

distinct forms of reasoning redundancy across 15 domains, utilizing the prompt provided in Figure 12. As shown in Table 7, TwT demonstrates exceptional efficiency, successfully resolving over 94% of the redundant steps observed in the SFT-with-RL baseline described in Section 3.2. Notably, it achieves a 97.2% resolution rate for Low-Density Long Descriptions, verifying its ability to compress verbose reasoning into high-density insights, while maintaining 87.3% resolution for structural issues like Over-Segmentation. These results confirm that the model has internalized resource-rational adaptive reasoning behavior, effectively activating System 2 reasoning for Rich Points while avoiding unnecessary elaboration on straightforward segments.

6.3 Translation Difficulty Adaptation

To better understand how models adapt their reasoning behavior to translation difficulty, we group the in-domain test set into three difficulty levels (Easy, Medium, Hard) estimated by DeepSeek-V3 (prompt in Figure 11). Table 8 reports the average quality and the response length for each group, averaged across all domains. Results indicate that TwT effectively addresses the reasoning redundancy of General-CoT through resource-rational allocation. On Easy inputs, it reduces token usage by 33% while improving quality by +6.4 points; conversely, on Hard inputs, it focuses on performance, achieving a substantial +6.7 points quality gain. Moreover, TwT models consistently outperform all SOTA LRMs across all difficulty levels in terms of quality, while maintaining significantly shorter reasoning traces—reducing average token usage by 32% compared to OpenAI-o3-mini and by 60% compared to Gemini-2.0-Flash-Thinking. This highlights TwT’s ability to generate concise, difficulty-aware reason-

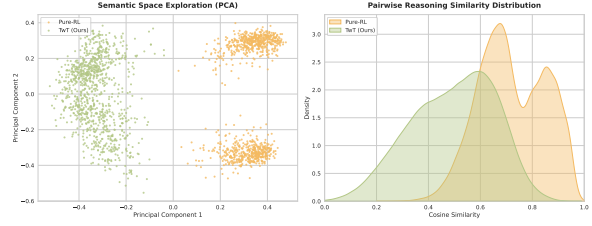


Figure 4: CoT trace similarity comparison. Pure RL vs TwT in space (Left) and distribution (Right).

ing while reducing overthinking.

6.4 Reasoning Collapse Mitigation

In Section 3.1, we identified a critical failure mode of pure RL, where reasoning rapidly degenerates into shallow, repetitive templates. To verify whether TwT successfully mitigates this reasoning collapse, we conduct a semantic diversity analysis across 15 domains. We compute pairwise cosine similarities of generated reasoning traces using multilingual Sentence-BERT⁵. Pure RL exhibits severe redundancy with a mean similarity of 0.89, whereas TwT significantly reduces this metric to 0.51. This divergence is visually corroborated by Figure 4, where the PCA projection (Left) shows pure RL confined to tight, isolated clusters compared to the broad semantic manifold of TwT, and the similarity histogram (Right) confirms that TwT diffuses the sharp redundancy peak of the baseline into a balanced distribution. These results show that TwT overcomes template dependency and encourages genuine reasoning.

6.5 Further Analysis

Appendix F provides additional analyses of TwT, including MQM error types, KL ablation, training dynamics, domain-aware prompting, inference cost, language consistency, and failure cases.

7 Conclusion

In this work, we present TwT, a resource-rational translation model that adapts reasoning effort to input difficulty. TwT combines difficulty-aware SFT and hybrid-reward RL to balance System 1 and System 2 behavior. Evaluated across diverse domains and languages, TwT matches or surpasses SOTA LRMs while reducing token usage by 32–60%, validating the effectiveness of aligning translation with human reasoning economy.

⁵sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Limitations

While **TwT** achieves robust performance across multiple domains, several limitations remain. First, the RL training data is randomly sampled without controlling for difficulty distribution, which may result in an imbalanced mix of easy, medium, and hard inputs. Second, the reasoning traces distilled from proprietary LLMs (e.g., DeepSeek-R1) may carry over implicit biases or domain preferences inherent in those models. Although our current setup yields consistent improvements, such biases could influence the reasoning behavior or stylistic tendencies of **TwT**. Finally, our current reward design does not incorporate difficulty-aware reward shaping. In particular, no length-based reward is applied to encourage concise reasoning on simple inputs and more detailed analysis for complex ones. Incorporating such adaptive rewards may further enhance the model’s ability to adjust reasoning depth based on input complexity in MDMT. We leave this direction for future work.

Acknowledgment

This work is supported by the National Science and Technology Major Project (Grant No. 2022ZD0116101), the National Natural Science Foundation of China (NSFC) under Grant No. 62206295, the Major Scientific Research Project of the State Language Commission in the 13th Five-Year Plan (Grant No. WT135-38), the public technology service platform project of Xiamen City (No. 3502Z20231043). In addition, we used a large language model to assist in polishing the visualizations in Figure 1 and generating certain decorative visual elements in Figure 3.

References

- M. Agar. 1994. *Language Shock: Understanding The Culture Of Conversation*. HarperCollins.
- Roei Aharoni and Yoav Goldberg. 2020. [Unsupervised domain clusters in pretrained language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.
- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). *Preprint*, arXiv:2402.17733.
- Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. [DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.
- Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and Min zhang. 2025. [Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis](#). *Preprint*, arXiv:2502.11544.
- Google DeepMind. 2024. [Introducing gemini 2.0: our new ai model for the agentic era](#). <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message>. Accessed: 2025-04-21.
- Google DeepMind. 2025. [Gemini 2.0 flash thinking](#). <https://ai.google.dev/gemini-api/docs/changelog>. Accessed: 2025-04-21.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Zhaopeng Feng, Shaosheng Cao, Jiahao Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025a. [Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning](#). *Preprint*, arXiv:2504.10160.
- Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025b. [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.
- Gemma Team, Morgane Riviere, and 1 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Daniel Gile and Victoria Lei. 2020. [Translation, effort and cognition](#). In *The Routledge handbook of translation and cognition*, pages 263–278. Routledge.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.

- 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, and 180 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- 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.
- Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, Hao Yang, Boxing Chen, and Osamu Yoshie. 2025. [R1-t1: Fully incentivizing translation capability in llms via reasoning learning](#). *Preprint*, arXiv:2502.19735.
- Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. [Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA. Association for Computational Linguistics.
- Kristian Tangsgaard Hvelplund. 2011. *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. Frederiksberg: Copenhagen Business School (CBS).
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helvar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Ifrimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, and 242 others. 2024. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. [Findings of the 2022 conference on machine translation \(WMT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23*, page 611–626, New York, NY, USA. Association for Computing Machinery.
- Isabel Lacruz. 2017. [Cognitive effort in translation, editing, and post-editing](#). *The handbook of translation and cognition*, pages 386–401.
- Zihao Li, Shaoxiong Ji, and Jörg Tiedemann. 2025. [Test-time scaling of reasoning models for machine translation](#). *Preprint*, arXiv:2510.06471.
- Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. [New trends for modern machine translation with large reasoning models](#). *Preprint*, arXiv:2503.10351.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2025a. [Introducing GPT-5](#). <https://openai.com/zh-Hans-CN/index/introducing-gpt-5/>.
- OpenAI. 2025b. [Openai o3-mini](#). <https://openai.com/index/openai-o3-mini/>. Accessed: 2025-04-21.
- 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.
- Qwen-Team. 2025. [Qwq-32b: Embracing the power of reinforcement learning](#).
- Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeira, Amin Farajian, and André F. T. Martins. 2025. [Tower+: Bridging generality and translation specialization in multilingual llms](#). *Preprint*, arXiv:2506.17080.
- 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.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T.

- Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. [Hybridflow: A flexible and efficient rlhf framework](#). In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys '25*, page 1279–1297, New York, NY, USA. Association for Computing Machinery.
- Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. [UM-corpus: A large English-Chinese parallel corpus for statistical machine translation](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. [Drt: Deep reasoning translation via long chain-of-thought](#). *Preprint*, arXiv:2412.17498.
- Jiaan Wang, Fandong Meng, and Jie Zhou. 2025b. [Deep reasoning translation via reinforcement learning](#). *Preprint*, arXiv:2504.10187.
- Jiaan Wang, Fandong Meng, and Jie Zhou. 2025c. [Extrans: Multilingual deep reasoning translation via exemplar-enhanced reinforcement learning](#). *Preprint*, arXiv:2505.12996.
- Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024a. [Findings of the WMT 2024 shared task on discourse-level literary translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics.
- Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. [Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.
- 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, Bangkok, Thailand. Association for Computational Linguistics.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. [A paradigm shift in machine translation: Boosting translation performance of large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. [Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation](#). In *Forty-first International Conference on Machine Learning*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, and Sitong Wang. 2025. [Ssr-zero: Simple self-rewarding reinforcement learning for machine translation](#). *Preprint*, arXiv:2505.16637.
- 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, Miami, Florida, USA. Association for Computational Linguistics.
- Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. [Demystifying long chain-of-thought reasoning in llms](#). *Preprint*, arXiv:2502.03373.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. [Dapo: An open-source llm reinforcement learning system at scale](#). *Preprint*, arXiv:2503.14476.
- 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, and Zheyang Luo. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A GRPO Algorithm

GRPO (Shao et al., 2024) extends PPO (Schulman et al., 2017) by removing the dependency on a value model and instead leveraging group-wise relative rewards estimation among sampled responses for more stable and efficient policy updates. Given a query \mathbf{x} , the model samples a group of G responses $\{\mathbf{y}_i\}_{i=1}^G$, each scored with reward r_i . The normalized advantage for each sample is computed as:

$$A_i = \frac{r_i - \text{mean}(\{r\}_{j=1}^G)}{\text{std}(\{r\}_{j=1}^G)}. \quad (1)$$

Then GRPO optimizes the policy model π_θ by maximizing the following objective:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = & \mathbb{E}_{\mathbf{x} \sim \mathcal{D}, \{\mathbf{y}_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | \mathbf{x})} \\ & \left(\frac{1}{G} \sum_{i=1}^G \min \left(\frac{\pi_\theta(\mathbf{y}_i | \mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}_i | \mathbf{x})} A_i, \right. \right. \\ & \left. \left. \text{clip} \left(\frac{\pi_\theta(\mathbf{y}_i | \mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}_i | \mathbf{x})}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) \right. \\ & \left. - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right), \end{aligned} \quad (2)$$

where $\pi_{\theta_{\text{old}}}$ and π_θ are the old and current policies, ϵ is the PPO clipping threshold, and β controls the weight of the KL regularization.

B Datasets

B.1 Details of Cold Start Data

For cold-start SFT, we collect a curated dataset of about 7K difficulty-adaptive Long CoT examples spanning 10 domains and three major translation directions: De→En, En→Zh, and Zh→En. The data is constructed via domain-aware generation with DeepSeek-R1 followed by difficulty-adaptive rewriting with GPT-4o (see Section 4.1). Detailed dataset statistics are presented in Figure 6.

B.2 Details of RL Training Data

We collect a diverse MDMT dataset for RL training across languages and domains. Specifically, we sample from the following sources:

Template

A conversation between User and Assistant. The user asks a translation question, and the Assistant solves it. The Assistant first thinks about the translation reasoning process in the mind, and then provides the final translation. The translation reasoning process and the final translation are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> translation reasoning process here </think> <answer> final translation here </answer>.\n\nUser: {Translation question}.\nAssistant: <think>

Figure 5: Template for pure RL in MT task.

- The German-English multi-domain dataset (Aharoni and Goldberg, 2020), including five distinct domains: IT, Law, Medical, Koran, and Subtitles.
- The English-Chinese UM-Corpus (Tian et al., 2014), covering four domains: News, Laws, Subtitles, and Science.
- The Chinese-English GuoFeng-Webnovel dataset (Wang et al., 2023, 2024a) from WMT23 and WMT24 literary translation tasks, representing the Literary domain.

For each domain, we randomly select 2K sentence pairs with a minimum source sentence length of 20 words (characters for Chinese) to ensure meaningful reasoning potential. This results in a total of 20K training samples used for RL training.

B.3 In-Domain Test Data

For in-domain evaluation, we use the official test sets associated with the corpora in Appendix B.2. For the Literary domain, we merge the valid_1, valid_2, test_1, and test_2 subsets to form a comprehensive test set. The data statistics is illustrated in Table 9.

B.4 Out-of-Domain Test Data

For out-of-domain evaluation, we consider a diverse set of test sets spanning multiple language pairs and domains. Specifically, the Conversation, Ecommerce, and Social domains are drawn from the WMT22 shared tasks (Kocmi et al.,

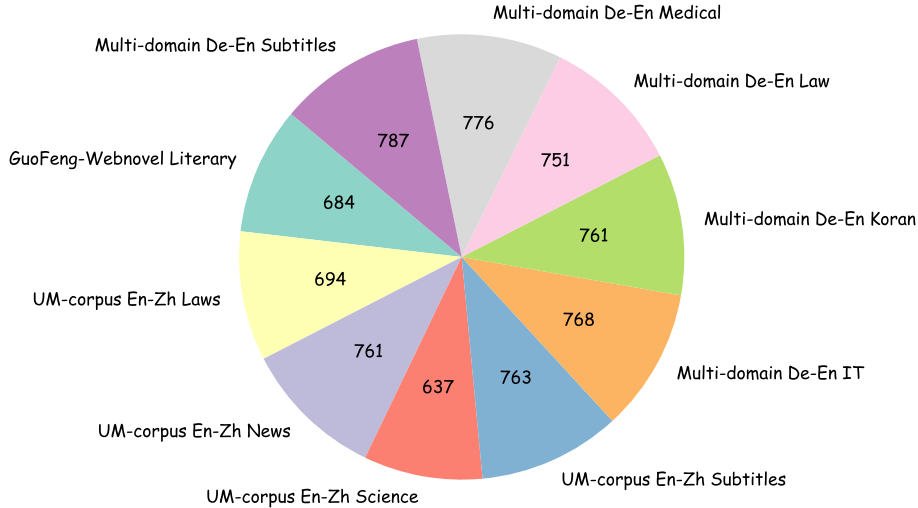


Figure 6: Distribution of our curated difficulty-adaptive Long CoT data.

Domain	Num	Domain	Num
En→Zh Laws	456	De→En IT	2000
En→Zh Subtitles	597	De→En Koran	2000
En→Zh Science	503	De→En Medical	2000
En→Zh News	1500	De→En Law	2000
Zh→En Literary	3038	De→En Subtitles	2000

Table 9: In-domain test sets and the number of samples for En↔Zh and De→En translation tasks.

Domain	Num	Domain	Num
En→Zh Conversation	484	En→Zh Social	511
En→Zh Ecommerce	530	En→Zh Culture	778
Zh→En CommonSense	1200	Zh→En Conversation	349
Zh→En Social	491	Zh→En Ecommerce	518
De→En Conversation	462	De→En Social	515
De→En Ecommerce	501		

Table 10: Out-of-domain test sets and sample counts for En↔Zh and De→En translation tasks.

2022). The Culture domain is sourced from the CAMT dataset (Yao et al., 2024), and the CommonSense domain comes from the CommonMT benchmark (He et al., 2020). The data statistics is illustrated in Table 10.

B.5 Domain Diversity Design

To ensure comprehensive evaluation and robust generalization of TwT across diverse translation scenarios, we designed the dataset to reflect a broad range of linguistic and contextual complexity, including: high-resource domains (e.g., News); low-resource scenarios (e.g., Koranic texts); terminology-heavy fields (e.g., IT, Law, Medicine, Science, E-commerce); context-sensitive domains

(e.g., Culture, CommonSense); stylistically demanding content (e.g., Literature); noisier or informal genres (e.g., Subtitles, Conversation, Social media). This diversity ensures comprehensive evaluation across domain specificity, reasoning difficulty, and stylistic variation.

B.6 Multilingual Test Data

For unseen-language evaluation, we adopt the FLORES+ benchmark (NLLB Team et al., 2024) and construct an *unseen* language set to minimize evaluation leakage from the languages already involved in our baseline training data (Table 5). Concretely, we first exclude all languages that appear in the baseline training coverage, including Chinese (zh), English (en), German (de), French (fr), Spanish (es), Portuguese (pt), Italian (it), Russian (ru), Korean (ko), Dutch (nl), Czech (cs), Icelandic (is), Ukrainian (uk), Hindi (hi), Japanese (ja), Polish (pl), Swedish (sv), Hungarian (hu), Romanian (ro), Danish (da), Norwegian (no), and Finnish (fi). This step ensures that the En↔X results reflect generalization to genuinely unseen languages rather than memorization of language-specific post-training signals. Next, to enable consistent computation of COMET and CometKiwi across all unseen directions, we further restrict the remaining FLORES+ languages to those supported by our COMET/CometKiwi scorers. After these two filters, we obtain a final unseen-language set consisting of 59 languages (Table 12).

The *seen* languages—German (de), English (en), and Chinese (zh)—are used across all baseline training sets, and are covered by the follow-

Model	Accuracy (\downarrow)	Style (\downarrow)	Fluency (\downarrow)	Terminology (\downarrow)	Non-translation (\downarrow)
DeepSeek-V3	50.55	30.50	6.95	11.90	0.09
Gemini-2.0-Flash	52.55	29.55	6.91	10.81	0.18
GPT-4o	50.52	30.76	7.11	11.40	0.21
DeepSeek-R1	50.13	30.97	6.71	12.10	0.09
Gemini-2.0-Flash-Thinking	52.31	29.70	6.54	11.31	0.13
OpenAI-o3-mini	49.82	30.70	7.13	12.23	0.11
OpenAI-o1	50.68	29.69	7.12	12.37	0.14
TwT-Qwen2.5-14B-Instruct	55.02	26.37	6.89	11.64	0.08

Table 11: MQM-based error analysis across different LLMs and LRMs.

ode	Language	Code	Language	Code	Language
afr	Afrikaans	als	Albanian	amh	Amharic
asm	Assamese	bel	Belarusian	ben	Bengali
bos	Bosnian	bul	Bulgarian	cat	Catalan
cym	Welsh	ekk	Estonian	ell	Greek
epo	Esperanto	eus	Basque	fil	Filipino
gle	Irish	glg	Galician	guj	Gujarati
hau	Hausa	heb	Hebrew	hrv	Croatian
hye	Armenian	ind	Indonesian	jav	Javanese
kan	Kannada	kat	Georgian	kaz	Kazakh
khk	Mongolian	khm	Khmer	kir	Kyrgyz
lao	Lao	lit	Lithuanian	lvs	Latvian
mal	Malayalam	mar	Marathi	mkd	Macedonian
mya	Burmese	npi	Nepali	pan	Punjabi
pbt	Pashto	plt	Malagasy	san	Sanskrit
sin	Sinhala	slk	Slovak	slv	Slovenian
som	Somali	srp	Serbian	sun	Sundanese
swh	Swahili	tam	Tamil	tel	Telugu
tha	Thai	tur	Turkish	uig	Uyghur
urd	Urdu	vie	Vietnamese	xho	Xhosa
ydd	Yiddish	zsm	Malay		

Table 12: The 59 unseen languages $\mathcal{L}_{\text{unseen}}$ used for En \leftrightarrow X evaluation after filtering FLORES+ by (i) post-training language coverage of evaluated backbones and (ii) COMET/CometKiwi language support.

ing datasets: the German-English multi-domain dataset (Aharoni and Goldberg, 2020), the English-Chinese UM-Corpus (Tian et al., 2014), and the Chinese-English GuoFeng-Webnovel dataset.

C Training Data Scale of MT Baselines

To ensure a fair comparison, we report the training data size used by each MT-specialized baseline. While TwT and most reasoning-augmented models, including MT-R1-Zero-7B, CoT-FT-7B, SFT-Parallel, and mExTrans-7B, are trained on approximately 27K examples, several other models leverage significantly larger corpora. For instance, TowerInstruct series is trained on 637K examples, Tower-Plus-9B on 286K, and ALMA-R on 21K. SSR-X-Zero-7B is trained on a notably smaller subset of 13K instances.

D API Details

The following APIs were used to access the SOTA LLMs and LRMs evaluated in our experiments:

- **OpenAI:** gpt-4o-2024-11-20, o1-2024-12-17, o3-mini-2025-01-31, and gpt-5-2025-08-07
- **DeepSeek:** deepseek-chat-2024-12-26, and deepseek-reasoner
- **Gemini:** gemini-2.0-flash, and gemini-2.0-flash-thinking-exp-01-21

E Experimental Details for Preliminary Analysis

E.1 Setup for Pure RL

Training Data. We construct a diverse MDMT training data with 20K samples by randomly sampling 2K sentence pairs from 10 domains, spanning three major language pairs: De \rightarrow En (Aharoni and Goldberg, 2020), En \rightarrow Zh (Tian et al., 2014), and Zh \rightarrow En (Wang et al., 2023, 2024a). We select only examples with source sentences longer than 20 words (characters for Chinese) to encourage meaningful reasoning. Evaluation is conducted using the test sets provided by the original datasets.

Training Template. To guide the base model toward producing translation-relevant reasoning behavior, we modify the template used in DeepSeek-R1-Zero (Guo et al., 2025) to better fit the translation task. The chat template is shown in Figure 5.

Reward Design. We follow DeepSeek-R1 (Guo et al., 2025) in using two reward types.

Format Reward: We apply regex matching to check whether the model responses adhere to the specified format. The format reward r_f is set to 1 if the format is correct, and -1 otherwise.

Quality Reward: We experiment with several quality metrics to evaluate the final translation output. Specifically, we consider BLEU⁶ (Papineni

⁶<https://github.com/mjpost/sacrebleu>

et al., 2002), a metric based on n-gram lexical overlap, and COMET⁷ (Rei et al., 2020) and CometKiwi⁸ (Rei et al., 2022), neural metrics that estimate semantic similarity using contextual embeddings through reference-based or reference-free scoring. All metric scores are normalized to the range $[0, 1]$ before being used in the reward function. During training, the quality reward r_q is computed based on one or more of these metrics. It is only applied when the output format is correct; otherwise, we assign a fixed penalty of $r_q = -2$. The final reward used during RL training is the sum of the format and quality components: $r = r_f + r_q$.

RL Optimization. We train the model using GRPO algorithm (Shao et al., 2024), which improves stability over PPO (Schulman et al., 2017) by leveraging group-wise relative advantages rather than explicit value functions. See Appendix A for full formulation.

Model Training. We conduct experiments using Qwen2.5-7B-Instruct (Yang et al., 2024) as the base model, and train it for 1 epoch on our collected 20k MDMT data. To investigate the effectiveness of different reward signals, we experiment with five quality reward variants: BLEU, COMET, CometKiwi, BLEU+COMET, and BLEU+CometKiwi.

E.2 Setup for RL with SFT

Training Data. We distill reasoning traces from the DeepSeek-R1 model (Guo et al., 2025) and apply additional quality filtering, resulting in about 7K high-quality Long CoT examples for MDMT.

Model Training. We fine-tune Qwen2.5-7B-Instruct (Yang et al., 2024) on the constructed Long CoT dataset for 1 epoch, followed by RL training on the 20K multi-domain data for 1 epoch, using a composite reward of BLEU and COMET. General-CoT and Domain-CoT share the same training pipeline and hyperparameters, differing only in the prompt used for SFT data construction: General-CoT uses "Translate the following {src_lang} text into {tgt_lang}.", while Domain-CoT uses "Translate the following {src_lang} text into {tgt_lang} while maintaining the domain style of the source text."

⁷Unbabel/wmt22-comet-da

⁸Unbabel/wmt22-cometkiwi-da

β	BLEU	COMET	KIWI	Quality	Lens.
0	30.07	82.35	81.20	64.54	130
0.001	30.58	82.52	81.21	64.77	278
0.005	30.79	82.48	81.28	64.85	273
0.01	30.37	82.49	81.40	64.75	282
0.02	30.24	82.37	81.29	64.63	254
0.05	29.71	82.23	81.24	64.39	275

Table 13: In-domain results with different KL coefficient.

F Further Analysis

F.1 Error-Type Analysis under the MQM Framework

To further investigate what types of translation errors reasoning helps address, we conducted a detailed analysis based on the MQM (Multidimensional Quality Metrics) framework. For error classification, we used an external LLM as an annotator, excluding "Source Error" as it pertains to low-quality source data. Given the distinct challenges of MDMT compared to general MT, we extended error types such as "Style Inconsistency", "Cross-domain Confusion", and "Terminology Misuse" to better capture the advantages of reasoning in MDMT. As shown in Table 11, reasoning-enhanced LRMs lead to lower error rates in Style, Fluency, Terminology, and Non-translation Errors. We believe this is due to the structured reasoning process, which contributes to a more human-like translation process and better error reflection, in contrast to the pattern-matching of traditional bilingual fine-tuning. Notably, after incorporating domain-aware CoT data, TwT reduced style errors by 3.18%, significantly improving the model's style adaptation. This highlights the effectiveness and applicability of our approach in MDMT.

F.2 KL Coefficient Analysis.

Recent study (Yu et al., 2025) suggests that removing KL regularization can enhance Long CoT reasoning. To examine whether this holds in the MDMT setting, we investigate the effect of the KL coefficient β in Table 13. However, our findings diverge from this conclusion: setting β to 0—completely removing the KL constraint—results in significantly shorter and less informative outputs. In contrast, small non-zero values (e.g., $\beta = 0.001$ or 0.005) achieve the best overall quality scores while maintaining reasonable response lengths. Larger values (e.g., $\beta = 0.02$ or 0.05) overly constrain the

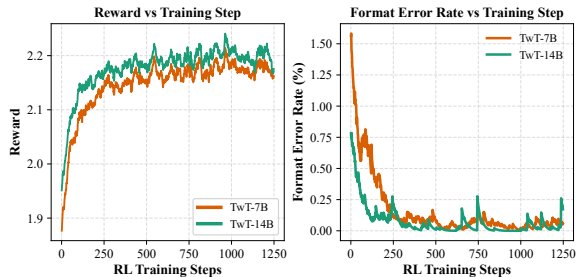


Figure 7: Average reward and format error rate over RL training steps.

policy, slightly degrading translation quality and increasing length instability. These results reflect the importance of balancing generation stability and flexibility: moderate KL regularization helps suppress degenerate behavior while preserving adaptive, difficulty-aware reasoning traces.

F.3 Analysis of Training Dynamics.

To better understand how RL shapes TwT’s reasoning and translation behavior, we analyze the training dynamics of TwT-7B and TwT-14B from two complementary perspectives.

Analysis of Reward and Format Stability Figure 7 further illustrates the reinforcement learning dynamics of TwT-7B and TwT-14B in terms of average reward and format error rate. Both models show a rapid increase in reward within the first few hundred steps, followed by a stable plateau, indicating that the hybrid reward formulation enables efficient convergence toward high-quality translation reasoning. The larger TwT-14B model consistently achieves higher rewards throughout training, suggesting stronger optimization capacity and better utilization of the reward signal. In contrast, the format error rate decreases sharply during the initial phase and remains near zero thereafter, demonstrating that the format reward effectively reduces structural inconsistencies and output anomalies as training progresses.

Translation Quality and Reasoning Depth Figure 8 shows the training dynamics of TwT-7B and TwT-14B during RL. BLEU, COMET, and CometKiwi scores steadily improve, indicating that our hybrid reward effectively enhances translation quality. Interestingly, response length increases in the early and mid stages, reflecting deeper reasoning, but decreases in later steps as the model learns to maintain quality with more concise traces. This indicates convergence to efficient, difficulty-

Table 14: General prompt vs. domain-aware prompt across domains.

Prompt	IT			Law			Medical		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
General prompt	36.66	83.49	78.68	39.16	84.97	82.35	40.23	83.82	81.81
Domain-aware prompt	36.55	83.57	79.82	40.57	85.32	83.54	41.14	84.06	82.97

Table 15: SFT trained on general CoT vs. domain-aware CoT.

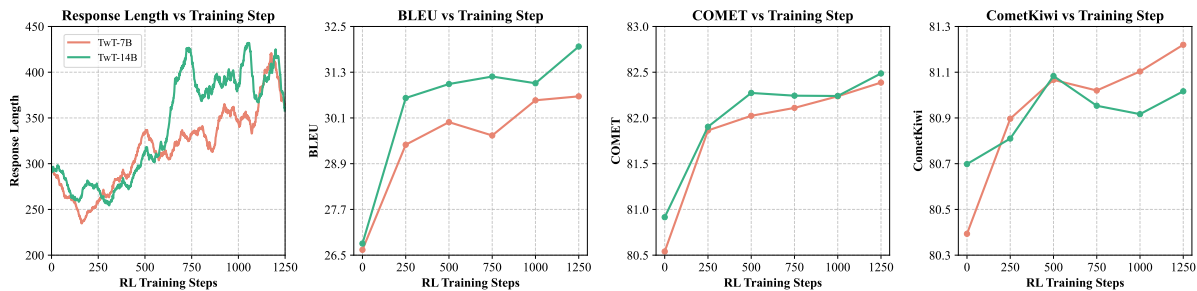
Method	In-Domain			Out-of-Domain		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI
SFT w/ general CoT	23.51	80.52	80.15	25.12	83.40	80.81
SFT w/ domain-aware CoT	23.74	80.74	80.46	25.67	83.57	81.18
Δ	+0.23	+0.22	+0.31	+0.55	+0.17	+0.37

aware reasoning behavior, as shown in Figure 8(a). Figure 8(b) and Figure 8(c) show the performance of TwT-7B and TwT-14B under different difficulty levels. In both models, BLEU, COMET, and CometKiwi scores are highest on easy inputs and lowest on hard ones, and response length increases with difficulty, indicating effective control of reasoning depth. Notably, TwT-14B generates shorter responses than TwT-7B on easy inputs, but longer responses on medium and hard inputs. This suggests that the larger model more effectively adapts its reasoning length to input difficulty, providing concise outputs when possible and allocating more reasoning to harder cases.

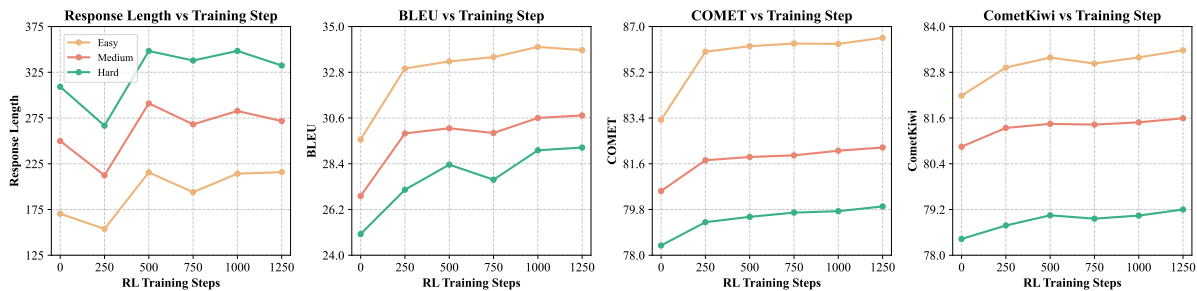
F.4 Analysis of Domain-Aware Prompting

To better understand the effect of domain-aware prompting, we conduct two complementary experiments that evaluate its impact from both the inference and training perspectives.

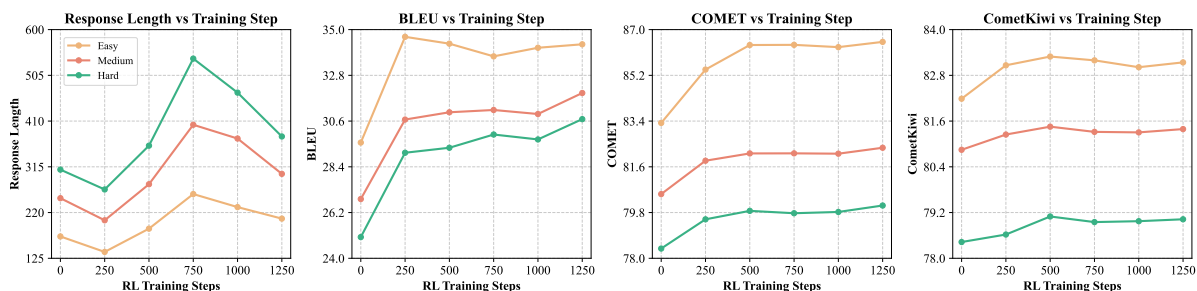
First, we investigate how prompt formulation affects translation quality when the model remains fixed (DeepSeek-R1). Two prompts are compared: a general prompt ("Translate the following src_lang sentence into tgt_lang.") and a domain-aware prompt ("Translate the following src_lang text into tgt_lang while maintaining the domain style of the source text."). As shown in Table 14, the comparison across three representative domains (IT, Law, Medical) demonstrates that domain-aware prompt explicitly instructs the model to infer and preserve domain-specific stylistic and terminological features. This better activates the reasoning capacity and leads to more accurate domain-aware translation. These findings motivated us to adopt domain-aware prompts throughout TwT’s training data curation phase, allowing domain signals to be explicitly injected dur-



(a) Training dynamics for TwT-7B and TwT-14B.



(b) Training dynamics for TwT-7B under different difficulty level.



(c) Training dynamics for TwT-14B under different difficulty level.

Figure 8: Training dynamics for TwT models.

ing learning.

Second, we evaluate the effect of using different prompts to construct Long-CoT data for SFT. Specifically, one dataset is generated using the general prompt, while another uses the domain-aware prompt to elicit domain-specific reasoning traces. As shown in Table 15, domain-aware CoT data lead to consistent improvements across all automatic evaluation metrics in both in-domain and out-of-domain scenarios. These results highlight that incorporating domain context enhances the model’s reasoning robustness and cross-domain generalization.

F.5 Computational Cost Analysis

To assess the computational efficiency of our reasoning-based translation paradigm, we compare the average inference time per example across dif-

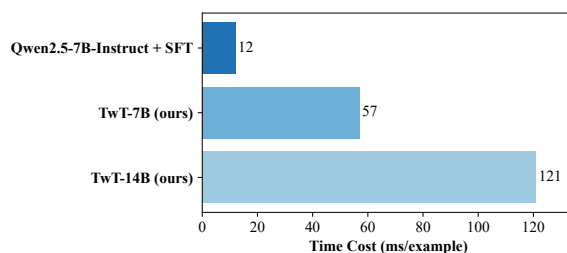


Figure 9: Inference time comparison.

ferent models. As shown in Figure 9, the reasoning-augmented TwT models incur additional computational overhead compared with the non-reasoning baseline (Qwen2.5-7B-Instruct + SFT). Specifically, TwT-7B and TwT-14B require 57 ms and 121 ms per example, respectively, compared to 12 ms for the baseline. Despite the higher inference cost, the time overhead remains acceptable given the

significant improvements in translation quality, reasoning accuracy, and style adaptation demonstrated in previous sections.

F.6 Language Consistency Discussion

To verify that **TwT**'s reasoning traces are linguistically coherent and aligned with the target translation language, we conducted a systematic analysis of language consistency throughout the data curation and training stages. We employed a language identification tool (`fastText`) to detect potential cross-lingual inconsistencies in reasoning traces, such as mixing English reasoning with Chinese or German summaries. In the initial R1-distilled CoT data, 9.26% (684/7382) of the reasoning traces contained incorrect languages. After GPT-4o rewriting, the inconsistency dropped sharply to 3.78% (279/7382). After SFT, the inconsistency rate on the test set further decreased to 0.04% (7/16094), and it reached 0% after the RL stage.

F.7 Reliability of Difficulty Estimation

To mitigate potential bias from relying solely on GPT-4o, we validated its difficulty estimation against two other strong models (DeepSeek-V3.2, Gemini-2.5-Flash) and human experts on 100 randomly sampled instances. Results demonstrate robust consistency across three dimensions: (1) **Internal Stability:** All models exhibited high self-agreement across 5 independent trials (GPT-4o: 0.92, DeepSeek: 0.93, Gemini: 0.91). (2) **Cross-Model Agreement:** The three models reached a 90% consensus rate after majority voting. (3) **Human Alignment:** Crucially, GPT-4o achieved the highest correlation with professional translators (**0.86**), surpassing DeepSeek (0.77) and Gemini (0.79). These findings confirm that GPT-4o provides a reliable and stable proxy for human judgment in our difficulty-aware framework.

F.8 Structured Analysis of Failure Cases

To provide a more systematic analysis, we further examine **TwT**'s bad cases across all domains. Specifically, we select the 10 lowest-quality examples from each of the 15 domains (150 cases in total) and analyze their CoT trajectories. We identify two recurring error patterns.

Reasoning–prediction misalignment. In some cases, the model makes the correct translation decision in its reasoning, but the final output fails to realize it. That is, the reasoning identifies an appropriate

Human Evaluation	TwT Win	TwT Lose	Tie
TwT-Qwen2.5-14B-Instruct vs. DeepSeek-R1 (671B)	0.41	0.47	0.12
TwT-Qwen2.5-14B-Instruct vs. DeepSeek-V3 (671B)	0.275	0.405	0.32
TwT-Qwen2.5-14B-Instruct vs. TowerInstruct-13B-v0.1	0.54	0.30	0.16

Table 16: Summary of human evaluation results for **TwT** compared with other models.

lexical choice or phrasing, while the prediction deviates from it. This suggests a gap between reasoning and realization. A possible remedy is to introduce an additional reward penalty during GRPO to explicitly discourage divergence between the decision expressed in the `<think>` block and the output in the `<answer>` block.

Domain–terminology misalignment. Another common failure occurs when the model correctly identifies the domain but does not switch to the appropriate terminology system. In such cases, the reasoning remains at a declarative level without consistently applying domain-specific lexical choices, syntactic preferences, fixed translations, or formatting conventions. One possible remedy is to incorporate external terminology lexicons or style-template libraries to guide domain-specific realization.

G Human Evaluation

G.1 Human Evaluation Analysis

We conducted a human evaluation to complement our automatic metrics. Specifically, we randomly sampled 100 examples from the Zh↔En test set, selecting 20 sentences from each of five domains. For each instance, the source sentence and two system outputs—one from **TwT-14B** and one from a comparison model—were independently evaluated by three professional translators, who selected the better translation or marked a tie when the quality difference was negligible. Table 16 summarizes the human evaluation results across three baselines: DeepSeek-R1, DeepSeek-V3, and TowerInstruct-13B-v0.1. These results show that while **TwT-14B** slightly underperforms DeepSeek-R1 and DeepSeek-V3—expected given its much smaller size—it still achieves a strong degree of parity. Importantly, **TwT-14B** outperforms TowerInstruct-13B-v0.1 in more than half of the evaluated examples (54% win rate), supporting the effectiveness of its reasoning-driven design in translation.

G.2 Comparison with Human Translation Reasoning

To further evaluate the alignment between TwT's reasoning process and that of professional human translators, we conducted a case study using a complex narrative sentence from a fictional novel, as shown in Figure 17. The source sentence features multiple subordinate and concessive clauses, complex logical progression, and dense world-specific terminology—a typical example in the fictional novel domain. Such sentences require precise control of syntax, consistent terminology management, and sensitivity to logical flow and pragmatic tone.

Similarities. TwT demonstrates several key reasoning behaviors consistent with professional translators. **First**, it correctly identified the sentence structure and logical relations, decomposing the original into two conceptual layers: (1) background explanation of the connection between the 神属大世界 and the 战星联邦, and (2) a reasoning-based denial of the assumed link between the 组织 and 神属大世界. This decomposition mirrors the syntactic and logical analysis stage in human translation reasoning, reflecting TwT's strong ability to detect hierarchical structure and causal relationships. **Second**, TwT shows strong terminological consistency, accurately translating key domain-specific entities such as "Immortal Holy Water" (不老圣水), "Divine Realm Major World" (神属大世界), and "War Star Federation" (战星联邦). While minor stylistic improvements remain possible, the model's use of consistent and semantically clear terms aligns with the terminology standardization step in professional translation workflows. TwT also handled subtle pragmatic expressions such as "有交情" and "换取" with contextually appropriate equivalents ("have connections with", "trade for"), indicating an emerging awareness of functional-pragmatic equivalence—a key element in expressing communicative intent and contextual tone. **Moreover**, before translating, TwT demonstrated an initial assessment of text difficulty and style, recognizing that the sentence belongs to a fictional novel with specialized terms. This awareness parallels a human translator's pre-translation difficulty assessment and background analysis, where genre and register are evaluated to inform strategy. TwT chose to preserve the source-style transliteration rather than domesticate the names, maintaining consistency with the fantasy universe's internal logic—an appropriate decision for this genre. **At the structural level**,

TwT effectively reconstructed the concessive–causal logic of "虽然...但...所以..." into the English pattern "Although...some...therefore...", faithfully capturing the original logical progression. This demonstrates the model's ability to rebuild syntactic and logical relations during target-language reorganization, consistent with human translators' reasoning in structural decomposition and coherence reconstruction. **In lexical judgment**, TwT made contextually informed choices, such as rendering "不老圣水" as "Immortal Holy Water" rather than the freer "elixir of youth" or "anti-aging water", thereby preserving the source's mythological tone. Similarly, translating "牵强" as "far-fetched" appropriately conveys the intended skepticism while maintaining stylistic naturalness. These decisions illustrate the model's ability to perform semantic disambiguation and lexical selection comparable to human translators' third-stage reasoning process. **Finally**, TwT's output features a natural syntactic flow and coherent discourse structure. For instance, "some powerful individuals...had connections with..." reconstructs the information hierarchy more fluently than a literal rendering would. This indicates an emerging sense of register and stylistic adaptation, partially fulfilling the requirements of the expression and style selection stage.

Differences. **First**, TwT's translation repeatedly uses the long-form expression such as "Divine Realm Major World", resulting in verbosity. Professional translators would typically balance terminological consistency with referential economy, adopting simplified references such as "the Divine Realm" or "that realm" to improve fluency and readability. For instance, a revised version could read: "Although 'Immortal Holy Water' is a specialty of the Divine Realm, some powerful individuals in the War Star Federation have connections with that realm and can obtain it through trade." This alternative maintains precision while achieving smoother rhythm and reduced redundancy. **Second**, TwT's stylistic control remains limited. Although the translation is grammatically accurate, it lacks the conversational tone and personality expected in dialogue. Since the original sentence is dialogue from Hans expressing skepticism, a professional translator would employ a more natural, idiomatic style to reflect the speaker's voice and pragmatic intent, e.g., "Come on, you can't just assume they're connected to the Divine Realm over a bit of Holy Water—that's a stretch!" Such adaptation better re-

produces conversational tone and aligns with the narrative style of fantasy dialogue. **Third, TwT's** cultural adaptation is surface-level. While "far-fetched" effectively captures the core meaning of "牵强," a human translator might enrich cultural accessibility by appending a light explanatory note, such as "Immortal Holy Water, a sacred elixir rumored to halt aging in the Divine Realm." This hybrid "foreignization plus domestication" strategy balances cultural retention with reader comprehension, commonly seen in professional literary translation.

H Prompt

The template for pure RL (Section 3.1) is shown in Figure 5. The prompt for generating difficulty-adaptive Long CoT data is provided in Figure 10. The prompt employed for the quantitative evaluation of reasoning efficiency is presented in Figure 12. The prompt for evaluating translation difficulty is shown in Figure 11 and the detailed examples for translation difficulty evaluation are provided in Table 17.

I Detailed Results by Metric

For completeness, we provide the full breakdown of BLEU, COMET, and CometKiwi for the results reported in Table 3, Table 4, and Table 5. The corresponding detailed results are shown in Table 18, Table 19, and Table 20, respectively.

You are an experienced translation expert. Your task is to optimize a given initial translation thought process by simulating the thinking process of a professional human translator.

Human translators typically first assess the difficulty of the translation task. Based on this difficulty level, they allocate appropriate time and cognitive effort:

- For simple translation tasks, they provide an accurate translation quickly with concise thought process.
- For moderate translation tasks, they reason carefully through any ambiguities to produce an accurate translation, without excessive expansion.
- For difficult translation tasks, they conduct deep and comprehensive thinking, exploring different translation strategies, comparing and verifying options, and refining every detail to produce the best translation.

Based on this professional approach, please optimize the given translation thought process by:

- Integrating the assessment of translation difficulty naturally into the thought process.
- Adjusting the depth and style of reasoning based on the difficulty level.
- Preserving the original language style and tone of the initial translation thought process.

Finally, output the optimized thought process and the final translation in JSON format with the keys "thought" and "translation", as shown below:

```
{
  "thought": "Optimized translation thought process based on difficulty",
  "translation": "Final translation based on the corrected thought process"
}
```

Source text:

Target language:

Initial translation thought process:

Figure 10: Full prompt used for generating difficulty-adaptive Long CoT data with GPT-4o.

Source Sentence	Reference Translation	Difficulty	Reason
“我只是想要听听您的意见。”她用一副恭敬的口吻说道。	“I just want to hear your opinion,” she said in a respectful tone.	Easy	A simple sentence with everyday vocabulary and direct grammar. No specialized knowledge required.
At the same time the waves are fanning out, they are also separating by wavelength, a process known as dispersion.	当这些波向外扩散时，它们也在按波长分开，这一过程叫频散。	Medium	Contains a compound sentence and one scientific term (*dispersion*), requiring moderate structural and lexical control.
Die folgende Tabelle zeigt die Häufigkeit und Schwere unerwünschter Wirkungen, die in >5% von 168 Patienten mit Mesotheliom berichtet wurden...	The table below provides the frequency and severity of undesirable effects reported in >5% of 168 mesothelioma patients...	Hard	Long and syntactically complex, includes multiple subordinate clauses, verb-final word order, and domain-specific medical terminology. Requires structural parsing and domain expertise.

Table 17: Detailed examples for translation difficulty evaluation.

Your task is to assess the difficulty of translating a given {src_lang} sentence into {tgt_lang}. Please evaluate the difficulty based on the following criteria:

1. Sentence complexity: Determine if the sentence is a simple sentence, a compound sentence, or includes subordinate clauses and other complex structures.
2. Vocabulary difficulty: Assess whether the sentence contains commonly used words or specialized terms or slang.
3. Grammar differences: Analyze if the sentence's grammatical structure is similar to or differs significantly from {tgt_lang}.
4. Contextual understanding: Consider whether understanding specific cultural contexts or background knowledge is necessary for accurate translation.

The difficulty level should be rated as "easy", "medium", or "hard". Additionally, provide a brief, simple reason for the assigned difficulty level. Output the result in JSON format with the keys "level" and "reason", as follows.

```
{  
  "level": "easy/medium/hard",  
  "reason": "simple explanation of the difficulty level."  
}
```

Here is the {src_lang} sentence: {src_text}

Figure 11: Full prompt used for evaluating translation difficulty with GPT-4o.

You are an expert model specialized in “Chain-of-Thought (CoT) Quality Evaluation for Multi-Domain Translation Tasks.” Your task is to compare the reasoning traces of Model A and Model B: Identify redundancy types in Model A’s reasoning and determine whether Model B has successfully eliminated these redundancies.

Please strictly adhere to the following **Redundancy Type Definitions** for your analysis. If you identify a new redundancy type in Model A, you may include it in the output with a supplemental explanation.

[Multi-Domain Translation: Redundancy Definitions]

- **Over-segmentation:** Excessively decomposing simple sentences or obvious content into word-by-word or phrase-by-phrase fragments, causing unnecessary step inflation.
- **Unnecessary linguistic explanation:** Over-explaining common sense or obvious linguistic points (grammar, part-of-speech, etymology, etc.) that do not influence the translation outcome.
- **Semantic repetition:** Repeating the explanation of the same meaning or using different expressions to illustrate the same semantic point without adding value to the translation.
- **Irrelevant information:** Introducing background knowledge, domain trivia, speculative content, or information unrelated to translation decisions (e.g., unnecessary context assumptions).
- **Redundant alternative translations:** Providing multiple translation candidates with highly similar meanings and low contribution, or performing iterative optimization on trivial differences.
- **Low-density long descriptions:** Using verbose sentences to express simple content or using a large volume of text to derive obvious reasoning steps.

[Task Requirements]

- Identify redundant snippets in Model A’s CoT and label them with the corresponding redundancy type.
- Explain why the snippet belongs to this category.
- Determine if Model B’s CoT eliminated this specific redundancy (Yes/No).
- Output strictly as a structured JSON array in the following format:

```
[  
  {  
    "type": "Name of the redundancy type",  
    "before_snippet": "Text snippet from Model A",  
    "reason": "Reason for the judgment",  
    "after_resolved": "Yes/No"  
  },  
  ...  
]
```

Now, please analyze the following two reasoning chains:

Model A CoT: {model_a_cot}

Model B CoT: {model_b_cot}

Figure 12: The prompt used for the quantitative evaluation of reasoning efficiency.

Source Text:

Oh, ich kann nicht glauben, dass sie weggelaufen ist.

Long CoT data distilled from DeepSeek-R1 (562 tokens):

<think>Okay, let's tackle this translation task. The user wants the German sentence "Oh, ich kann nicht glauben, dass sie weggelaufen ist." translated into English. First, I need to understand the sentence. The main components are "Oh, I can't believe that she ran away." The speaker is expressing surprise or disbelief that someone (female) has run away. Next, the user mentioned identifying the domain and ensuring the translation reflects the domain's style. Hmm, domains could be things like literature, casual conversation, formal reports, etc. The original sentence uses "Oh" and "kann nicht glauben," which are pretty colloquial. It sounds like conversational or maybe even a dialogue from a novel or a TV show. The emotional tone here is informal and expressive. So the domain is likely casual or conversational. Maybe even a dramatic context, like a soap opera or a novel. Now, translating it to match that style. The direct translation would be "Oh, I can't believe she ran away." But depending on the domain, maybe we can make it a bit more natural. For example, in casual speech, people might contract "cannot" to "can't," which is already done. "Weggelaufen" is "ran away," which is correct. Wait, the original uses "sie," which could be "she" or "they" depending on context. But since the user provided the German sentence, and the verb is "weggelaufen ist" which is singular (since "ist" is singular), so "she" makes sense here. Unless the context is formal, but the sentence seems informal. So "she" is correct. Also, the structure "dass sie weggelaufen ist" is a subordinate clause. In English, "that she ran away." But in conversational English, sometimes people drop "that," so "Oh, I can't believe she ran away." That sounds more natural and fits the casual domain better. So the domain here is probably everyday conversation or a dramatic dialogue. The style should be informal, using contractions and natural phrasing. So the translation should reflect that. No need for formal language. I think that's it. Let me double-check. The original is expressing surprise, so the translation needs to keep that emotional tone. "Oh" at the beginning sets that. "Can't believe" is better than "cannot believe" for informality. Yes. So the final translation would be "Oh, I can't believe she ran away." And the domain is casual conversation or possibly a narrative in fiction. The style is informal and expressive.</think>

<answer>Oh, I can't believe she ran away.</answer>

Our curated difficulty-adaptive Long CoT data (167 tokens):

<think>First, I assess the difficulty of the translation: this is a simple, conversational German sentence expressing surprise or disbelief about someone's action. There are no idioms or complex structures, so the translation is straightforward. The sentence is "Oh, ich kann nicht glauben, dass sie weggelaufen ist."—literally, "Oh, I can't believe that she ran away." To preserve the informal and expressive tone, it's best to use the contraction "can't" and drop the explicit "that" for natural conversational English. "Oh" at the start effectively conveys the sense of surprise. No ambiguities or special cultural references require deeper reasoning. Thus, the optimal translation is clear and concise.</think>

<answer>Oh, I can't believe she ran away.</answer>

Source Text:

现在他要找的就是外门弟子所修炼的地方。

Long CoT data distilled from DeepSeek-R1 (476 tokens):

<think>好的，我现在要处理这个翻译任务。用户给的中文句子是：“现在他要找的就是外门弟子所修炼的地方。”首先，我需要确定这句话的领域。看起来像是武侠小说或者修仙类的小说，因为提到了“外门弟子”和“修炼的地方”，这些词汇常见于这类文学作品中。接下来，我需要确保翻译符合目标领域的风格。武侠或修仙小说通常使用较为正式且带有古风的英语表达，同时保留一些特定的术语。比如“外门弟子”通常翻译为“outer sect disciples”，而“修炼的地方”可以是“training grounds”或者“practice area”。这里可能需要根据上下文选择更合适的词汇，但“training grounds”听起来比较自然。然后，我需要分析句子的结构。原句的结构是“现在他要找的就是……的地方”，强调当前的目标。翻译时要保持这种强调，可能需要用“What he needed to find now was...”这样的结构，这样既符合英文表达习惯，又传达了原句的重点。另外，时态方面需要注意。原句中的“现在”可能指的是当前的情况，用过去时的话可能更合适叙述故事，所以用“needed”比较合适，因为很多小说是用过去时叙述的。不过如果是现在进行的情节，可能用现在时，但根据上下文，假设是过去时更合适。还要确保术语的一致性。比如“外门弟子”在整个作品中可能有固定的翻译，所以保持一致性很重要。如果用户之前有特定的译法，可能需要调整，但这里没有其他信息，所以采用通用的译法。最后，检查整个句子的流畅性和准确性。确保没有遗漏信息，并且符合目标领域的风格。比如是否需要用更古风的词汇，或者是否需要调整语序来更符合英语习惯。比如原句中的“所修炼的”可以用定语从句处理，放在“地方”后面，这样更自然。总结一下，翻译后的句子应该是：“What he needed to find now was the training grounds of the outer sect disciples.”这样既准确传达了原意，又符合武侠或修仙小说的风格。</think>

<answer>What he needed to find now was the training grounds of the outer sect disciples.</answer>

Our curated difficulty-adaptive Long CoT data (325 tokens):

<think>在接到这个翻译任务时，我首先要对原句“现在他要找的就是外门弟子所修炼的地方。”进行难度评估。从内容来看，这是一句来自武侠小说的叙述句，涉及专有名词“外门弟子”和“修炼的地方”，但整体语法结构和表达并不复杂。因此，我将其归为一般难度的文学文本，需要注意术语准确和风格匹配。具体处理步骤如下：
1. 理解背景：确认“外门弟子”为宗门中层级较低的弟子，常见翻译为“outer sect disciples”；“修炼的地方”可译为“the place where they cultivate”或更自然的“training grounds”。
2. 语法与时态：向首“现在”表明当前动作，但考虑到小说常用过去时，主句采用了“needed to find”，更符合英语叙述习惯。
3. 保持强调：原文强调“现在他要找的就是……”，英文中可通过“what he needed to find now was...”或倒装强调，也可简语表达为“Now, he needed to find the...”。
4. 风格：保持叙述的剧情感和正式度。
结合以上分析，最终优化的translation为“Now, he needed to find the place where the outer sect disciples cultivated.”这样既准确传递了原文语境和术语，也兼顾了英文的流畅和符合武侠小说风格。</think>

<answer>Now, he needed to find the place where the outer sect disciples cultivated.</answer>

Figure 13: Examples of our curated difficulty-adaptive Long CoT data (Easy and Medium).

Source Text:

During a meeting with chargé d'affaires ad interim Li Zhimin of the Chinese Embassy in Vietnam on 15 June 1956, Vice Foreign Minister of the Democratic Republic of Vietnam Ung Van Khiem solemnly stated that, "according to Vietnamese data, the Xisha Islands and Nansha Islands are historically part of Chinese territory." Le Loc, Acting Director of the Asian Department of the Vietnamese Foreign Ministry, who was present, specifically cited Vietnamese data and pointed out that, "judging from history, these islands were already part of China at the time of the Song Dynasty."

Long CoT data distilled from DeepSeek-R1 (648 tokens):

<think>好的，我需要把这段英文翻译成中文。首先，先通读一遍原文，理解内容。这段文字是关于1956年越南民主共和国副外长与中国驻越南使馆临时时代办会面的记录，涉及西沙和南沙群岛的主权归属问题。显然，这属于外交或历史领域的文本，可能涉及法律或国际关系方面的术语。

接下来，确定领域。这里涉及的是历史事件和外交声明，属于国际关系或历史文献的领域。因此，翻译时需要保持正式、严肃的语气，使用准确的官方术语，比如“临时代办”、“副外长”、“郑重声明”等。

然后处理具体的翻译点：

1. "chargé d'affaires ad interim" 正确翻译是“临时代办”，注意头衔的准确。
2. "solemnly stated" 要译为“郑重声明”，符合外交场合的正式用语。
3. "Vietnamese data" 翻译为“越南方面的材料”比较合适，避免直译成“数据”。
4. "Acting Director" 是“代理司长”。
5. "cited Vietnamese data" 这里可能需要译为“列举了越南方面的材料”，保持上下文一致。
6. "Song Dynasty" 是“宋朝”，注意朝代名称的正确翻译。

另外，要注意时间、职务和人名的正确翻译。比如“Ung Van Khiem”应译为“翁文谦”，“Le Loc”译为“黎禄”。需要确认这些越南人名的标准中文译名，可能需要查阅相关资料确保准确性。

在句式结构上，英文中的时间状语从句和直接引语的结构要转换为中文习惯的表达方式，比如将“During a meeting on 15 June 1956”译为“1956年6月15日，……会晤时”，并将直接引语用引号正确标出。

最后检查整个翻译是否流畅，是否符合外交文献的正式风格，术语是否一致，人名、地名、机构名是否准确无误。比如“亚洲司代理司长黎禄”是否正确，是否有遗漏的职务或头衔。

可能还需要注意一些细节，比如“西沙群岛”和“南沙群岛”的正确译名，以及“历史资料”和“根据历史来看”等表达是否准确传达原意。总结来说，整个过程需要准确理解原文内容，确定合适的领域和风格，选择合适的术语和句式，确保翻译的准确性和专业性。</think>

<answer>1956年6月15日，越南民主共和国外交部副部长翁文谦在与中国驻越南大使馆临时时代办李志民会晤时郑重声明：“根据越南方面的材料，西沙群岛和南沙群岛从历史上看属于中国领土。”在场的越南外交部亚洲司代理司长黎禄特别列举了越南方面的材料指出：“从历史上看，这些岛屿早在宋朝时期就已属中国。”</answer>

Our curated difficulty-adaptive Long CoT data (550 tokens):

<think>首先对翻译难度进行评估：本段原文内容涉及1950年代中越之间的外交历史事件，主体为正式外交辞令和归属表述，内容较为严肃，需要在译文中准确还原称谓、人名、地名，确保外交和官方用语的严谨性。人名、职务译名需要查证权威译法，同时还要处理引号、直接引语及时间状语的顺畅转化。因此，属于较高难度的历史外交领域翻译任务，需要分步细致推敲。

首先，通读原文，精准捕捉每个时间点、人物、头衔及关键论述。

1. "chargé d'affaires ad interim"标准译法为“临时代办”；"vice foreign minister"译为“副外长”，而越南民主共和国副外长通用译名为“翁文谦”；中国驻越南使馆临时代办“李志民”。
2. 时间状语“During a meeting ... on 15 June 1956”应译为“1956年6月15日，在中国驻越南使馆临时时代办李志民的会见中”。
3. 语气要郑重，外交辞令使用“郑重声明”。
4. "Vietnamese data"一般译为“越南方面资料”；
5. "acting director"常见译法为“代司长”，“Le Loc”应对照权威译名为“黎禄”。
6. "judging from history"——译为“从历史上看”或“据史可证”，结合上下文选用更习惯的表达。
7. 人名和地名要核查官方标准译名。

8. 引号方面，层层嵌套时需注意区分人物直述与转述，建议采用中文标准引号和冒号。译文顺序可适当调整以更贴近中文表达习惯。

完成初稿后，逐句核对语气严谨、表达完整，查漏补缺，确保术语统一和风格正式。再次复核关键历史用词、机构名称准确无误，并确保译文逻辑通顺。</think>

<answer>1956年6月15日，越南民主共和国外交部副部长翁文谦在会见中国驻越南使馆临时时代办李志民时郑重声明：“根据越南方面的资料，西沙群岛和南沙群岛自古就是中国的领土。”当时在场的越南外交部亚洲司代理司长黎禄还专门引用了越南方面的资料指出：“从历史上看，这些群岛在宋代时就已经属于中国。”</answer>

Figure 14: Examples of our curated difficulty-adaptive Long CoT data (Hard).



Prompt: Translate the following sentence from Chinese into English.

“我们回去！”



<think>**First**, I assess the difficulty of this translation. The source sentence “我们回去！” is short and clear, with no cultural nuances or domain-specific terms, making it a **simple task**.

Next, I analyze each component: “我们” translates directly as “we”; “回去” means “go back”, and in English, “Let's go back!” is accurate and natural. Considering possible contexts (such as **action movies, novels, or casual conversation**), the phrase should maintain its brevity and emphasis. Since there are no special cultural references or ambiguities, a concise and faithful translation is appropriate.

Overall, the optimal approach is to translate word-for-word accurately and preserve the original tone.</think>

<answer>“Let's go back!”</answer>

153 tokens



Prompt: Translate the following sentence from German into English.

Neuen Eintrag hinzufügen



<think>**First**, I assess the difficulty of translating “Neuen Eintrag hinzufügen” from German to English. The sentence consists of a straightforward phrase with no cultural nuances or domain-specific jargon, making it a **simple task**.

Next, I analyze each component: “Neuen” translates directly as “new”; “Eintrag” means “entry”; and “hinzufügen” is accurately rendered as “add”. In **technical contexts, such as software interfaces**, precision and brevity are valued, but there is no need for elaborate strategies.

Reviewing possible alternatives like “Add new entry”, “Add new item”, or “Add entry”, I find that “Add new entry” maintains accuracy and fidelity to the original while fitting typical UI conventions in English.

Overall, this translation requires minimal cognitive effort beyond accurate word choice and attention to register.</think>

<answer>Add new entry</answer>

182 tokens

Figure 15: Output examples of our TWT-7B.



Prompt: Translate the following sentence from Chinese into English.

身穿皮甲的中年男子站在甲板旁，向下俯瞰着这一望无际的草原与美景。



<think>**First**, I assess the difficulty of this translation. The sentence involves **literary description** and requires attention to accuracy in character details and imagery, but there is no deep cultural nuance or technical challenge, **so I consider it a moderate task**.

Next, I analyze each component: '身穿皮甲的中年男子' translates as 'a middle-aged man wearing leather armor', accurately conveying both the attire and age. '站在甲板旁' becomes 'stood on the deck', which maintains fidelity and fits typical novelistic style. '向下俯瞰着这一望无际的草原与美景' means looking down at the endless grassland and beautiful scenery below; 'looked down at the vast grassland and the beautiful scenery below' preserves the original's emphasis and flow. To optimize, I ensure that 'vast grassland' is appropriately rendered as 'the endless prairie' for naturalness and accuracy.

Overall, the sentence should be concise yet vivid, suitable for an English-language novel.

After reviewing possible alternatives, such as 'a middle-aged man in leather armor stood on the deck, looking down at the endless grassland and the beautiful scenery below', I confirm that this translation maintains accuracy, fidelity, and appropriate tone.</think>

<answer>A middle-aged man in leather armor stood on the deck, looking down at the endless grassland and the beautiful scenery below.</answer>

278 tokens



Prompt: Translate the following sentence from German into English.

ii) von Rückständen dieser Stoffe in lebenden Tieren, deren festen und fluessigen Ausscheidungen sowie im Tiergewebe und in tierischen Erzeugnissen wie Fleisch, Milch, Eiern und Honig beschreiben;



<think>**First**, I assess the difficulty of this translation. The sentence involves precise **legal** terminology and requires accuracy in mapping each phrase to its appropriate English equivalent, but there is no unusual cultural nuance or ambiguity, making it **a moderate task**.

Next, I analyze each component: 'von Rückständen dieser Stoffe' translates as 'of residues of these substances', which is accurate and faithful to the source. 'In lebenden Tieren' becomes 'in live animals', a direct and correct choice. 'Ihre festen und fluessigen Ausscheidungen' refers to solid and liquid excreta, accurately rendered as 'their solid and liquid excreta'. 'Sowie im Tiergewebe' means 'and in animal tissues', with 'animal tissue' being an appropriate and concise term.

Finally, 'und in tierischen Erzeugnissen wie Fleisch, Milch, Eiern und Honig beschreiben' should become 'describe in animal products such as meat, milk, eggs and honey'; here, 'products' is suitable and maintains precision, though 'derived products' could also be considered if broader context demands it, but given the specificity of the list, 'animal products' is accurate and clear.

Overall, the translation should preserve accuracy and flow typical for formal regulatory language.</think>

<answer>(ii) the residues of such substances in live animals, their solid and liquid excreta, in animal tissues and in animal products such as meat, milk, eggs and honey;</answer>

310 tokens

Figure 16: Output examples of our TWT-7B.



Prompt: Translate the following sentence from Chinese into English.

“虽然‘不老圣水’是神属大世界的特产，但在战星联邦的一些强者与神属大世界的强者也有交情，也会换取一些‘不老圣水’，所以单是这点‘不老圣水’就认定组织与神属大世界有关，就有些牵强了！”汉斯摇头说道。



<think>**First**, I assess the difficulty of this translation. The source sentence involves domain-specific terms from **fictional novels**, such as '不老圣水', '神属大世界', and '战星联邦', requiring moderate to high accuracy in terminology and logical reasoning. Therefore, it is a **hard** task suitable for careful attention to accuracy and flow.

Next, I analyze each component: '不老圣水' should be rendered as 'Eternal Youth Holy Water'; '神属大世界' refers specifically to a 'Divine Realm Major World', which can be translated as 'Divine Realm World' or 'Divine Realm Realm', with both accurate and concise options available; '战星联邦' is accurately translated as 'War Star Federation'. For the phrase '在战星联邦的一些强者与神属大世界的强者也有交情，也会换取一些‘不老圣水’', I need to ensure that the logic flows smoothly in English, suggesting a connection between the powerful figures from the War Star Federation and those from the Divine Realm World, who might exchange some 'Eternal Youth Holy Water'. This could be expressed as 'Some powerful individuals in the War Star Federation have connections with the powerful beings in the Divine Realm World, and they would also trade for some "Eternal Youth Holy Water"', maintaining accuracy and fidelity.

Finally, '单是这点“不老圣水”就认定组织与神属大世界有关，就有些牵强了!' means 'Just because there is a little bit of "Eternal Youth Holy Water", it's too strong to conclude that the organization is related to the Divine Realm World!', emphasizing skepticism. A natural rendering would be: 'However, since "Immortal Holy Water" is a specialty of the Divine Realm World, but some powerful individuals in the War Star Federation also have connections with the powerful beings in the Divine Realm World, and they would also trade for some 'Immortal Holy Water', so it's a bit too far-fetched to conclude that the organization has anything to do with the Divine Realm World just based on this small amount of 'Immortal Holy Water'!

Overall, the translation requires precision and logical coherence, fitting the style of a novel.</think>
<answer>"Although the 'Immortal Holy Water' was a specialty of the Divine Realm Major World, some powerful individuals in the War Star Federation also had connections with the powerful beings in the Divine Realm Major World, and they would also trade for some 'Immortal Holy Water'. Therefore, it was a bit too far-fetched to conclude that the organization was related to the Divine Realm Major World just because of this small amount of 'Immortal Holy Water'!" Hans shook his head and said.</answer>

556 tokens

Figure 17: Output examples of our TwT-7B.

Method	Laws				News				Science				Subtitles			
	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens
<i>Large Language Models</i>																
DeepSeek-V3	59.58	<u>89.15</u>	84.42	-	<u>35.86</u>	86.94	85.44	-	33.54	88.28	85.46	-	23.66	82.44	82.69	-
Gemini-2.0-Flash	56.43	88.90	84.39	-	36.39	86.37	85.14	-	<u>33.47</u>	87.69	85.24	-	24.64	81.91	82.33	-
GPT-4o	48.50	88.28	84.52	-	34.58	86.22	85.02	-	32.09	87.34	84.84	-	24.91	81.91	81.93	-
<i>Large Reasoning Models</i>																
DeepSeek-R1	<u>59.34</u>	<u>89.15</u>	84.68	577	33.35	86.67	85.40	498	31.82	88.08	<u>85.47</u>	478	20.50	82.10	<u>82.84</u>	514
Gemini-2.0-Flash-Thinking	55.75	88.61	84.27	702	33.90	86.02	84.79	1149	31.98	87.47	85.06	1092	23.83	80.90	81.22	708
OpenAI-o3-mini	42.73	87.67	84.44	428	32.90	86.29	85.35	443	31.38	87.42	85.24	385	22.95	81.90	82.40	355
OpenAI-o1	48.11	88.32	<u>85.13</u>	478	33.93	86.61	<u>85.47</u>	408	32.32	87.92	85.56	367	22.96	82.31	82.60	340
GPT-5	54.06	89.34	85.43	784	35.05	86.65	85.52	740	32.21	88.06	<u>85.47</u>	606	23.23	82.32	82.68	519
QwQ-32B	43.77	87.35	84.39	667	33.24	85.77	84.63	584	32.05	87.11	84.42	563	22.58	81.54	81.29	584
<i>MT-Specialized Models</i>																
SFT-Parallel-7B	56.64	88.82	84.28	-	28.89	85.43	83.91	-	29.28	86.19	83.68	-	26.93	81.41	80.26	-
ALMA-7B-R	36.34	85.45	81.86	-	24.53	83.95	81.64	-	22.16	84.33	81.83	-	18.56	80.18	79.40	-
ALMA-13B-R	40.33	86.89	83.10	-	26.87	84.65	82.43	-	24.37	85.72	82.59	-	19.54	81.11	80.00	-
TowerInstruct-7B-v0.2	50.53	88.25	82.96	-	30.72	84.61	82.45	-	27.76	85.60	82.99	-	22.45	80.75	80.01	-
TowerInstruct-13B-v0.1	52.28	88.57	83.09	-	32.24	85.36	83.09	-	29.43	86.00	83.17	-	23.74	81.19	80.74	-
CoT-FT-7B	57.05	88.76	84.35	51	30.17	85.23	83.31	42	28.72	86.02	83.48	39	28.17	80.92	79.20	29
MT-R1-Zero-7B	35.49	86.78	84.51	72	31.69	86.01	84.54	64	29.45	86.87	84.61	61	22.48	81.73	81.70	55
SSR-X-Zero-7B	38.88	86.44	83.44	56	28.14	85.87	83.98	52	28.54	87.11	84.65	49	23.18	80.50	81.99	39
mExTrans-7B	38.77	87.11	84.44	597	25.05	86.07	85.34	553	25.74	87.34	85.31	546	14.31	81.40	82.85	476
<i>Our Models</i>																
TwT-Qwen2.5-7B-Instruct	52.56	88.76	84.73	310	33.63	<u>86.75</u>	84.88	311	32.41	87.53	84.77	294	24.74	<u>82.63</u>	81.73	247
TwT-Qwen2.5-14B-Instruct	56.03	89.07	84.65	320	34.57	86.65	84.63	285	32.75	87.54	84.66	272	24.44	82.73	81.74	241

Method	Literary				IT				Koran				Medical				Average			
	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens
<i>Large Language Models</i>																				
DeepSeek-V3	16.14	77.28	77.07	-	38.16	83.90	78.86	-	17.94	74.91	80.33	-	41.43	84.15	81.90	-	33.29	83.38	82.02	-
Gemini-2.0-Flash	<u>18.37</u>	77.20	76.67	-	37.93	83.20	78.49	-	19.70	74.97	79.71	-	44.39	84.49	81.76	-	<u>33.91</u>	83.09	81.72	-
GPT-4o	17.75	77.48	77.18	-	37.23	83.52	78.37	-	17.65	75.04	80.64	-	41.89	84.23	81.89	-	31.83	83.00	81.80	-
<i>Large Reasoning Models</i>																				
DeepSeek-R1	11.25	75.44	75.16	574	36.66	83.49	78.68	593	17.05	74.76	80.63	790	40.69	83.94	82.39	667	31.33	82.95	81.91	586
Gemini-2.0-Flash-Thinking	18.01	77.06	76.63	781	37.26	82.94	78.40	345	19.57	75.03	79.79	677	43.36	84.14	81.70	415	32.96	82.77	81.48	734
OpenAI-o3-mini	17.48	76.77	76.67	546	37.13	82.63	78.20	343	15.52	73.54	80.19	511	39.82	83.60	81.79	346	29.99	82.48	81.78	420
OpenAI-o1	16.54	77.50	77.81	521	36.40	83.20	<u>79.25</u>	403	16.76	75.03	82.02	506	40.44	83.88	83.29	441	30.93	83.10	82.70	433
GPT-5	15.01	76.72	77.08	859	36.92	83.58	<u>79.22</u>	492	18.83	<u>75.49</u>	<u>81.80</u>	751	43.00	84.29	<u>83.13</u>	531	32.29	83.31	<u>82.54</u>	660
QwQ-32B	12.99	75.98	76.31	863	21.93	82.95	79.43	583	12.36	73.59	81.47	963	33.95	83.27	82.82	735	26.61	82.20	81.85	693
<i>MT-Specialized Models</i>																				
SFT-Parallel-7B	15.77	76.73	75.37	-	<u>40.64</u>	<u>84.08</u>	79.35	-	21.09	74.66	78.39	-	43.76	84.40	82.47	-	32.88	82.72	80.96	-
ALMA-7B-R	13.46	75.12	74.97	-	33.75	81.12	77.72	-	14.09	71.92	79.12	-	37.97	83.04	81.36	-	25.11	80.64	79.74	-
ALMA-13B-R	14.20	75.89	75.81	-	34.22	81.63	77.78	-	14.75	72.76	79.63	-	40.34	83.38	81.49	-	26.83	81.50	80.35	-
TowerInstruct-7B-v0.2	15.54	75.49	74.93	-	38.20	83.47	78.67	-	10.99	69.02	70.13	-	<u>46.81</u>	<u>84.58</u>	80.80	-	30.38	81.47	79.12	-
TowerInstruct-13B-v0.1	16.81	76.01	75.46	-	39.26	83.85	78.44	-	11.41	69.23	69.16	-	48.55	84.99	80.93	-	31.72	81.90	79.26	-
CoT-FT-7B	15.31	76.45	74.96	52	40.41	83.79	79.27	35	20.52	74.04	78.16	45	44.21	84.34	82.59	46	33.07	82.44	80.66	42
MT-R1-Zero-7B	13.75	76.93	76.72	69	34.51	82.71	79.59	56	13.05	72.84	80.66	71	27.28	83.31	82.69	71	25.96	82.15	81.88	65
SSR-X-Zero-7B	13.66	76.82	76.14	54	27.01	80.08	76.57	36	13.97	72.86	79.75	46	28.86	82.12	80.98	50	25.28	81.47	80.94	48
mExTrans-7B	10.43	76.16	76.62	610	25.54	78.76	77.19	452	11.80	73.88	80.56	604	25.40	81.66	81.93	565	22.13	81.55	81.78	551
<i>Our Models</i>																				
TwT-Qwen2.5-7B-Instruct	17.93	<u>78.35</u>	<u>77.30</u>	281	40.12	84.04	80.05	222	21.13	75.45	80.19	269	43.74	84.32	82.81	262	33.28	<u>83.48</u>	82.06	274
TwT-Qwen2.5-14B-Instruct	19.10	78.40	76.93	354	41.26	84.17	79.68	234	22.20	75.71	80.13	336	44.45	84.42	82.89	287	34.35	83.59	81.91	291

Table 18: Detailed metric breakdown for Table 3. We report in-domain translation results across eight domains, averaged over En→Zh, Zh→En, and De→En.

Method	Conversation				Ecommerce				Social			
	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens
<i>Large Language Models</i>												
DeepSeek-V3	36.76	87.03	81.48	-	32.29	85.53	81.19	-	32.26	84.59	81.44	-
Gemini-2.0-Flash	38.05	86.90	81.37	-	<u>32.41</u>	85.49	80.87	-	32.98	84.28	81.04	-
GPT-4o	<u>38.02</u>	86.79	81.43	-	32.88	<u>85.57</u>	81.04	-	<u>32.75</u>	84.47	81.08	-
<i>Large Reasoning Models</i>												
DeepSeek-R1	33.34	86.47	81.37	534	27.41	85.03	80.91	552	27.22	83.94	81.17	554
Gemini-2.0-Flash-Thinking	37.21	86.64	81.24	1204	31.50	85.24	80.74	822	31.79	83.89	80.83	1081
OpenAI-o3-mini	36.66	86.31	81.15	290	31.51	85.13	80.97	363	31.79	84.00	80.96	372
OpenAI-o1	35.50	<u>87.10</u>	<u>82.81</u>	327	29.65	85.61	82.13	399	29.21	<u>84.50</u>	82.43	405
GPT-5	35.25	87.14	82.82	448	29.09	85.37	<u>82.02</u>	609	29.00	84.10	<u>82.31</u>	652
<i>MT-Specialized Models</i>												
SFT-Parallel-7B	32.20	84.11	80.62	-	26.94	82.92	80.12	-	25.88	81.61	80.12	-
ALMA-7B-R	29.49	84.67	79.96	-	25.13	82.98	79.50	-	25.76	82.51	79.77	-
ALMA-13B-R	31.86	85.48	80.76	-	26.24	83.75	80.12	-	27.00	83.06	80.47	-
CoT-FT-7B	31.58	84.09	80.63	31	26.71	82.95	80.11	45	24.98	81.37	79.83	42
MT-R1-Zero-7B	32.38	85.88	81.50	53	26.98	84.25	81.25	66	26.39	83.28	81.40	65
SSR-X-Zero-7B	30.65	85.32	81.13	37	25.64	83.94	80.79	50	25.47	83.19	81.26	49
mExTrans-7B	24.53	84.72	81.15	464	20.46	83.52	81.24	566	19.60	82.29	81.70	555
<i>Our Models</i>												
TwT-Qwen2.5-7B-Instruct	35.10	86.37	81.69	231	30.92	84.85	81.35	273	30.79	84.08	81.58	269
TwT-Qwen2.5-14B-Instruct	35.35	86.39	81.58	240	31.04	85.11	81.21	309	31.04	83.91	81.46	298

Method	Culture				CommonSense				Average			
	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens
<i>Large Language Models</i>												
DeepSeek-V3	40.27	85.46	83.21	-	32.55	85.36	79.98	-	<u>34.83</u>	85.59	81.46	-
Gemini-2.0-Flash	<u>39.11</u>	85.02	82.94	-	31.60	84.71	79.38	-	<u>34.83</u>	85.28	81.12	-
GPT-4o	38.65	85.24	83.14	-	<u>32.54</u>	<u>85.26</u>	79.87	-	34.97	<u>85.47</u>	81.31	-
<i>Large Reasoning Models</i>												
DeepSeek-R1	35.88	<u>85.45</u>	83.14	560	28.50	84.52	79.99	602	30.47	85.08	81.32	561
Gemini-2.0-Flash-Thinking	36.94	84.63	83.68	1220	32.15	84.50	79.68	2335	33.92	84.98	81.23	1332
OpenAI-o3-mini	34.05	84.36	84.25	596	28.23	84.28	<u>80.85</u>	436	32.45	84.82	81.64	411
OpenAI-o1	34.57	85.06	84.25	542	28.43	84.97	80.95	392	31.47	85.45	82.52	413
GPT-5	35.79	85.04	<u>84.15</u>	984	27.08	84.43	80.75	530	31.24	85.22	<u>82.41</u>	645
<i>MT-Specialized Models</i>												
SFT-Parallel-7B	31.81	82.87	81.35	-	22.19	81.91	78.98	-	27.81	82.68	80.24	-
ALMA-7B-R	33.25	83.81	82.83	-	23.25	82.80	80.00	-	27.38	83.35	80.41	-
ALMA-13B-R	22.23	81.12	79.48	-	27.35	82.41	78.97	-	26.94	83.16	79.96	-
CoT-FT-7B	30.41	82.46	81.10	54	22.38	81.97	78.88	33	27.21	82.57	80.11	41
MT-R1-Zero-7B	31.86	83.99	82.85	79	24.50	82.94	79.53	51	28.42	84.07	81.31	63
SSR-X-Zero-7B	28.70	83.25	81.11	66	23.93	82.82	79.80	34	26.88	83.70	80.82	47
mExTrans-7B	27.12	84.30	83.92	631	18.61	81.92	79.23	470	22.06	83.35	81.45	537
<i>Our Models</i>												
TwT-Qwen2.5-7B-Instruct	35.24	84.59	83.64	352	29.08	84.30	80.07	219	32.23	84.84	81.67	269
TwT-Qwen2.5-14B-Instruct	37.70	84.71	83.23	333	29.81	84.36	79.98	259	32.99	84.90	81.49	288

Table 19: Detailed metric breakdown for Table 4. We report OOD translation results across five domains, averaged over En→Zh, Zh→En, and De→En.

Method	En→Zh				Zh→En				De→En			
	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens
<i>Large Language Models</i>												
Qwen2.5-7B-Instruct	35.05	85.44	83.11	-	12.63	69.88	70.22	-	28.82	79.36	80.53	-
Gemma-2-9B-IT	32.66	84.49	82.36	-	13.57	72.41	71.94	-	25.84	77.06	79.01	-
<i>MT-Specialized Models</i>												
ALMA-7B-R	26.69	84.13	81.43	-	13.46	75.12	74.97	-	28.87	79.66	80.13	-
Tower-Plus-9B	37.91	86.76	<u>84.88</u>	-	16.56	77.88	77.78	-	<u>36.56</u>	82.21	82.36	-
SFT-Parallel-7B	38.58	86.34	83.39	-	15.77	76.73	75.37	-	35.72	81.63	81.11	-
MT-R1-Zero-7B	32.00	86.23	84.34	62	13.75	76.93	76.72	69	25.56	80.39	81.73	66
SSR-X-Zero-7B	34.20	<u>86.86</u>	84.58	50	14.83	77.45	77.27	54	26.65	80.39	81.55	45
mExTrans-7B	28.63	86.44	85.02	537	10.43	76.16	76.62	610	22.06	79.27	80.88	544
<i>Our Models</i>												
TwT-Qwen2.5-7B-Instruct	<u>38.28</u>	87.27	84.43	298	<u>17.93</u>	78.35	<u>77.30</u>	281	35.53	81.93	81.90	256
TwT-Gemma-2-9B-IT	36.34	86.65	84.23	227	19.17	<u>78.11</u>	<u>77.07</u>	249	36.73	<u>82.15</u>	<u>81.97</u>	218

Method	En→X				X→En				Average			
	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens	BLEU	COMET	KIWI	Tokens
<i>Large Language Models</i>												
Qwen2.5-7B-Instruct	5.97	55.56	51.30	-	20.45	76.22	74.30	-	20.58	73.29	71.89	-
Gemma-2-9B-IT	14.05	75.63	<u>73.16</u>	-	<u>31.55</u>	<u>84.71</u>	<u>82.88</u>	-	23.53	<u>78.86</u>	77.87	-
<i>MT-Specialized Models</i>												
ALMA-7B-R	2.78	58.13	73.13	-	12.01	63.94	58.17	-	16.76	72.20	73.57	-
Tower-Plus-9B	6.07	63.86	69.46	-	27.99	81.38	79.68	-	<u>25.02</u>	78.42	<u>78.83</u>	-
SFT-Parallel-7B	1.19	52.16	43.97	-	18.47	75.58	73.05	-	21.95	74.49	71.38	-
MT-R1-Zero-7B	5.91	58.76	58.27	358	21.91	77.48	75.61	74	19.82	75.96	75.34	126
SSR-X-Zero-7B	6.02	58.94	56.34	306	20.73	78.05	75.78	48	20.49	76.34	75.10	101
mExTrans-7B	5.09	61.33	61.74	1047	15.95	76.80	75.21	731	16.43	76.00	75.89	694
<i>Our Models</i>												
TwT-Qwen2.5-7B-Instruct	5.99	59.57	57.83	483	22.94	77.73	75.84	328	24.13	76.97	75.46	329
TwT-Gemma-2-9B-IT	<u>11.12</u>	<u>72.02</u>	80.82	280	32.79	85.02	83.30	257	27.23	80.79	81.48	246

Table 20: Detailed metric breakdown for Table 5. We report results on seen and unseen language directions. En, Zh, and De are *seen* languages, while X denotes *unseen* languages; En→X and X→En report averages over English↔unseen-language directions.