Learning to Translate by Translating: Stabilizing the Dual Loop via Semantic-Aware Self-Evolution

Kui Liu[1]*, Mingming Yin[2], Zuoli Tang[1], Zihao Li[1], Chilin Fu[2], Xiaolu Zhang[2], Jun Zhou[2], Lixin Zou[1], Chenliang Li[1]†

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, [2]Ant Group

{liukui, tangzuoli, zihao.li, zoulixin, cllee}@whu.edu.cn

{yinmingming.ymm, chilin.fcl}@antgroup.com

{yueyin.zxl, jun.zhoujun}@antfin.com

Abstract

Despite the remarkable success of Large Language Models (LLMs) in Machine Translation (MT), the scarcity of high-quality parallel corpora and the prohibitive cost of their acquisition constrain scalability. To this end, we propose **L**earning to **T**ranslate by **T**ranslating (**LTT**), an LLM-driven dual-learning framework that enables autonomous translation, achieving an 80.42% performance improvement over the base model. By adapting the cycle-consistency principle to the generative paradigm, LTT eliminates the need for parallel data. It employs a robust semantic-aware reward function that balances adequacy with reconstruction fidelity, effectively mitigating the reward hacking issues inherent in traditional unsupervised MT. Relying solely on monolingual data, our 8B model consistently outperforms significantly larger models (70B+) in low-resource settings and achieves parity with state-of-the-art supervised baselines on mainstream benchmarks. LTT thus offers a scalable, data-efficient paradigm for autonomous machine translation.

1 Introduction

LLM-based translation systems, such as Tower (Alves et al.) and X-ALMA (Xu et al., b), have achieved state-of-the-art (SOTA). However, this success is largely attributed to the effectiveness of supervised training on vast, human-curated parallel corpora. The high cost of corpus construction poses a major barrier to further improvement. To break the dependency on parallel corpora, leveraging monolingual data has long been a long-standing goal in MT research. Pioneering works in Dual Learning (He et al., 2016) and Unsupervised Machine Translation (UMT) (Lample et al., 2018) introduced the concept of utilizing round-trip consistency as a training signal. Despite their theoretical appeal, these early attempts struggled with instability and were prone to "reward hacking"—where models optimize for lexical reconstruction at the expense of semantic fidelity (Marchisio et al., 2020). Consequently, achieving robust, fully autonomous translation without references remains an open challenge.

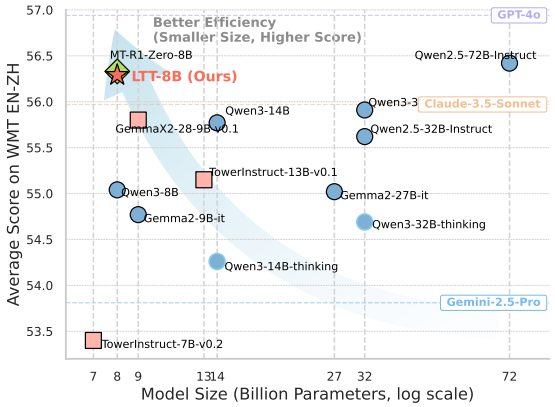

Figure 1: Overall performance of LTT-8B on the WMT EN-ZH benchmark. Our enhanced 8B model (red star) rivals the performance of models 4x-9x larger (e.g., Qwen2.5-72B) and GPT-4o, while outperforming all competitors within the <10B parameter class.

The advent of LLMs with reasoning and self-correction capabilities, such as OpenAI o1 (Jaech et al., 2024) and DeepSeek R1 (Guo et al., 2025), offers a unique opportunity to revisit and modernize this traditional framework. However, a naive integration is insufficient. As revealed in our analysis (Section 5.4), directly applying traditional reconstruction objectives to LLMs leads to catastrophic shortcuts, such as copying source text to maximize overlap scores. Furthermore, recent RL-based MT approaches, such as MT-R1-Zero (Feng et al., 2025), still rely heavily on ground-truth references for reward computation, limiting their scalability in low-resource scenarios.

To this end, we introduce LTT, an enhanced

*Work done during an internship at Ant Group

†Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 14711–14726

dual learning framework tailored for the LLM era. Unlike prior methods constrained by unstable optimization or simplistic metrics, LTT leverages Group Relative Policy Optimization (GRPO) for stable training and integrates a semantic-aware reward mechanism to prevent reward hacking. This allows the model to self-evolve using solely monolingual data, achieving performance comparable to, or even exceeding, supervised models.

Extensive experimental results demonstrate that LTT achieves parity with state-of-the-art baselines without relying on parallel corpora fine-tuning. On EN-ZH, our 8B model nearly matches optimal supervised benchmarks (lagging by only 0.07%) and surpasses proprietary systems like Gemini-2.5-Pro (77.58 vs. 77.55) in semantic evaluation. Notably, it outperforms the significantly larger Qwen2.5-72B-Instruct (76.52), highlighting its efficiency (see Figure 1). In challenging multilingual settings, LTT boosts the base model from 32.22 to 58.51, outperforming competitors like Gemma2-9B-it and even LLaMA-3.1-70B. Further analysis attributes these gains to our hybrid reward function, which synergizes lexical and semantic signals to mitigate reward hacking and ensure stable convergence.

Our contributions are summarized as follows:

- We propose LTT, an LLM-driven dual-learning framework that integrates GRPO into a self-evaluation loop, enabling effective self-evolution using solely monolingual data.
- We design a semantic-aware reward architecture that balances reconstruction accuracy with anti-cheating constraints, effectively mitigating the reward hacking problem inherent in unsupervised objectives.
- Extensive experiments show our 8B model matches supervised baselines and significantly outperforms much larger LLMs (e.g., 70B+) in low-resource settings.

2 Related Work

Machine Translation with Large Language Models. Cutting-edge machine translation with LLMs follows two main paradigms: in-context learning (ICL) and supervised fine-tuning (Anil et al., 2023; Gao et al., 2024; Li et al., 2024; Xu et al., a). By presenting few-shot demonstrations to LLMs, ICL circumvents the heavy computational costs of fine-tuning (Zhu et al., 2024), though at the expense of prompt sensitivity and performance instability (Agrawal et al., 2023). Fine-tuning, in contrast, improves the MT performance via supervised training on large-scale parallel datasets (Cui et al., 2025; Costa-Jussà et al., 2022), exemplified by representative models such as Tower and X-ALMA (Alves et al.; Guo et al., 2024; Xu et al., b). However, the rules of scaling laws constrain further performance improvements of fine-tuning. Moreover, both two paradigms exhibit a significant reliance on large-scale, human-annotated data, which limits scalability, particularly in low-resource scenarios.

Reasoning and Reinforcement Learning for Machine Translation. To transcend the limitations of supervised training, researchers have adopted RL, initially to mitigate exposure bias by optimizing global metrics like BLEU (Ranzato et al., 2016; Wu et al., 2017). Inspired by LLM reasoning, recent approaches incorporate CoT prompting to decompose translation into intermediate steps for enhanced fidelity (Feng et al., 2024; Wang et al., 2024). However, their success heavily relies on validating the reasoning process, often necessitating manually engineered templates, complex search algorithms like MCTS (Zhao et al., 2024), or external evaluators (He et al., 2025). Crucially, even leading RL frameworks such as MT-R1-Zero (Feng et al., 2025) and DeepTrans (Wang et al., 2025) still depend on reference translations for reward computation, restricting their potential for fully autonomous learning.

Dual Learning and Self-Evolution. Foundational research in dual learning (He et al., 2016) introduced a closed-loop feedback system leveraging the symmetry of translation tasks, while unsupervised MT (Lample et al., 2018) utilized shared latent spaces and back-translation for alignment. Recently, this paradigm has evolved into self-rewarding mechanisms for LLMs (Yuan et al., 2024; Zou et al., 2025). Despite their promise, these approaches remain fragile: early dual methods suffer from severe instability under domain mismatch (Marchisio et al., 2020), and generative self-rewarding models are prone to reward hacking, where they prioritize high internal scores over semantic fidelity (Wu et al., 2024). LTT revitalizes this framework by integrating GRPO with semantic-aware rewards. By replacing unstable policy gradients with GRPO and enforcing round-trip consistency through semantic metrics, our approach effectively mitigates semantic drift and reward hacking, enabling stable, autonomous self-

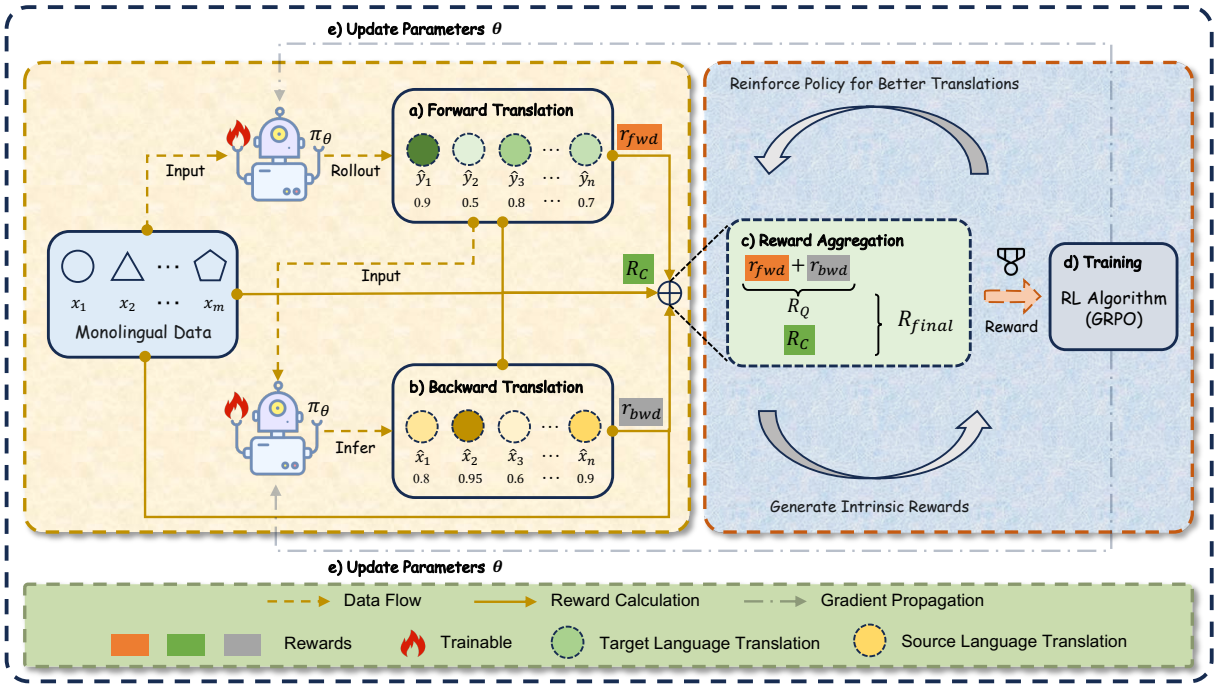

Figure 2: Overview of LTT. The actor model participates in both forward and backward translation. We consider reward signals from quality and anti-cheating, two perspectives, to update the model via GRPO.

improvement without parallel data.

3 Methodology

3.1 LTT Framework

Inspired by the primal-dual structure of traditional Dual Learning (He et al., 2016), our bidirectional translation loop leverages the pre-trained capabilities of LLMs and stabilizes the optimization process using GRPO with a semantic-aware reward composition. By recognizing round-trip consistency as an intrinsic reward signal, the system can self-evolve and improve within a closed-loop process (see Figure 2).

Specifically, we collect a batch of source sentences $\{x_i\}_{i=1}^{m}$ from language $\mathcal{L}_s$. The actor model, parameterized by a policy π_θ, performs two sequential steps:

1. Forward Translation ($\mathcal{L}_s \to \mathcal{L}_t$). The model generates a candidate $\hat{y}_i$ as the target language $\mathcal{L}_t$ translation:

$$\hat{y}_i \sim \pi_\theta(\cdot \mid x_i, \text{prompt}_{s\to t}).$$

As no reference translation is involved, this step necessitates a reference-free evaluation.

2. Backward Translation ($\mathcal{L}_t \to \mathcal{L}_s$). The process runs in reverse by translating $\hat{y}_i$ back into the source language $\mathcal{L}_s$ for $\hat{x}_i$ reconstruction:

$$\hat{x}_i \sim \pi_\theta(\cdot \mid \hat{y}_i, \text{prompt}_{t\to s}).$$

In this step, the original source sentence x_i serves as a high-quality, self-reference against which we can evaluate the reconstruction quality.

Unified Prompting Strategy. To ensure consistency, we employ a unified prompting template for both translation directions, following Feng et al. (2025). The model is instructed to place its final translation within `<translate>` tags and any intermediate reasoning within `<think>` tags. The full prompt details are available in Appendix A.1.

3.2 Reward Architecture

The success of LTT hinges on a tailored reward function that guides the model towards high-quality translations while preventing reward hacking. The final reward, R_{final}, is structured hierarchically to prioritize valid formatting before assessing translation quality:

$$R_{\text{final}}(x, \hat{y}, \hat{x}) = \begin{cases} 1 + R_{\text{Q}} + R_{\text{C}}, & \text{correct format,} \\ -3, & \text{otherwise.} \end{cases}$$

A penalty of -3 is assigned if the output $\hat{y}$ does not adhere to the required `<translate>` tag format. Since GRPO computes advantages via group-level normalization, the *relative ranking* of the penalty matters more than its absolute magnitude; any sufficiently negative value ensures malformed outputs receive the lowest advantage (see Appendix C for a sensitivity analysis). For valid outputs, we evaluate

both translation quality and anti-cheating capability, and additionally incorporate a base reward of $+1$.

3.2.1 Quality Reward ($R_{\mathbf{Q}}$)

This reward integrates two complementary metrics for a holistic assessment of translation quality, defined as: $R_{\text{Q}} = r_{\text{fwd}} + r_{\text{bwd}}$:

- **Forward Semantic Adequacy (r_{fwd}):** For the forward pass ($x \rightarrow \hat{y}$), we use **COMETkiwi**, a widely used reference-free metric. It evaluates the semantic adequacy of the translation by comparing the source and the hypothesis straightforwardly, formalized as $r_{\text{fwd}} = \text{COMETkiwi}(x, \hat{y})$.
- **Backward Reconstruction Fidelity (r_{bwd}):** For the back-translation ($\hat{y} \rightarrow \hat{x}$), we leverage the original source x as a reference. We use **BLEU** to measure the fidelity of the reconstruction, $r_{\text{bwd}} = \text{BLEU}(\hat{x}, x)$. A high score signifies that the target translation $\hat{y}$ preserved sufficient information to accurately recover the original input.

3.2.2 Anti-Cheating Reward ($R_{\mathbf{C}}$)

A vanilla self-supervised reward is fragile and unstable. To ensure robust learning and mitigate reward hacking, we introduce two penalty terms, $R_{\text{C}} = r_{\text{copy}} + r_{\text{mix}}$:

- **Source-Copying Penalty (r_{copy}):** A trivial failure mode is to copy the source ($x \approx \hat{y}$) to maximize the backward BLEU score. We address this by penalizing lexical overlap in the forward direction: $r_{\text{copy}} = -\text{BLEU}(x, \hat{y})$.
- **Language-Mixture Penalty (r_{mix}):** To discourage the generation of linguistically incoherent outputs, we apply a penalty of -0.5 if language mixing is detected in $\hat{y}$. This value acts as a soft threshold: once the penalty is sufficient to push mixed-language outputs to the bottom of the group ranking, further increases yield diminishing returns while risking over-penalization of valid entities such as untranslated acronyms (see Appendix C).

3.3 Policy Optimization with GRPO

We optimize our translation policy π_θ using GRPO (Shao et al., 2024; Guo et al., 2025). For each input, GRPO samples a group of G candidates from the current policy and computes a normalized advantage A_i for each based on its relative reward within the group. The policy is then updated by maximizing the standard GRPO objective:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{\hat{y}_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i(\theta)A_i, \text{clip}(\rho_i(\theta), 1-\varepsilon, 1+\varepsilon)A_i \right) - \beta D_{\text{KL}}\left(\pi_\theta(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)\right) \right], \quad (1)$$

where $\rho_i(\theta)$ is the probability ratio. This objective uses a clipped surrogate function and a KL-divergence penalty to ensure stable policy updates. Please refer to Appendix B.1 for further details.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments to evaluate the effectiveness of our LTT across bilingual and multilingual settings. Attribute to the merit of free-parallel-data in our methodology, we, thereby, select the popular parallel translation benchmarks and remove all target-side references for traning corpora construction.

- **Training Data:** For bilingual (EN-ZH) training, we source sentences from WMT 2017-2020 competitions, yielding $6,565$ sentences each for English and Chinese. For multilingual training, we create a diverse corpus from the FLORES-200 (Costa-Jussà et al., 2022) training set, covering EN/ZH paired with six other languages (DE, FR, ES, IT, JA, KO). We sample 500 pairs for each of the 24 translation combinations (e.g., EN→DE, ZH→DE), treating all sentences as monolingual, resulting in a final $12,000$-sentence corpus.
- **Evaluation Data:** We evaluate performance on test sets for fair comparison. For EN↔ZH, we use the official test sets from WMT23[1] and WMT24[2]. For multilingual tasks, we report results on the official FLORES-200 test set.

Evaluation Metrics. To ensure a comprehensive assessment of translation quality, we adopt a dual-metric approach that captures both lexical fidelity and semantic adequacy. We report case-sensitive **BLEU** scores computed via `sacrebleu` for standardized, reproducible measurement of n-gram overlap. To evaluate semantic preservation, we employ **xCOMET-XL** (Guerreiro et al., 2024),

[1]https://www2.statmt.org/wmt23/translation-task.html
[2]https://www2.statmt.org/wmt24/translation-task.html

Table 1: Main results on WMT and FLORES-200 benchmarks. The best and second-best scores are **bolded** and underlined, respectively. "†" indicates thinking mode. **Black model names** denote peers within the ∼7-9B parameter class, while *gray italics* represent larger models or closed-source systems included as references. XX denotes the average across all evaluated target languages for a given setting (i.e., DE, FR, ES, IT, JA, and KO for FLORES-200).

Models	WMT					FLORES-200								
	EN→ZH		ZH→EN		Avg.	EN→XX		XX→EN		ZH→XX		XX→ZH		Avg.
	BLEU	xCM	BLEU	xCM		BLEU	xCM	BLEU	xCM	BLEU	xCM	BLEU	xCM	
Closed Source														
Claude-3.5-Sonnet	38.21	75.54	22.95	87.16	55.97	32.76	92.69	34.48	97.00	21.26	91.19	37.39	84.01	61.35
GPT-4o	41.47	75.62	22.73	87.92	**56.94**	31.51	92.50	34.20	96.75	20.32	89.81	37.09	83.13	60.66
Gemini-2.5-Pro	32.28	77.55	19.80	85.63	53.81	33.14	95.05	33.14	96.56	22.25	92.21	36.51	87.27	**62.02**
Open Source														
General LLMs														
Qwen3-8B	36.56	74.94	22.67	85.98	55.04	25.47	88.98	31.44	94.75	17.23	85.21	32.92	77.19	56.65
Qwen3-8B†	26.97	67.31	16.71	80.11	47.77	22.54	87.67	27.28	91.18	15.20	83.28	33.43	78.31	54.86
Qwen3-14B	38.60	75.75	21.46	87.27	55.77	26.92	91.18	32.38	96.23	18.55	89.15	35.83	85.25	59.44
Qwen3-14B†	35.67	73.73	22.61	85.01	54.26	28.78	91.56	32.13	95.11	18.46	88.43	35.66	82.18	59.04
Qwen3-32B	39.37	75.44	21.52	87.31	55.91	30.53	92.69	34.25	96.50	19.61	89.33	37.11	85.38	60.67
Qwen3-32B†	38.37	74.55	20.20	85.65	54.69	24.61	91.79	30.41	95.00	14.94	88.25	33.04	82.51	57.57
Qwen2.5-32B-Instruct	39.28	75.16	21.19	86.87	55.62	28.04	90.62	32.29	96.26	18.01	87.67	36.01	83.91	59.10
Qwen2.5-72B-Instruct	40.02	76.52	21.88	87.27	56.42	30.48	92.39	34.83	96.78	19.46	89.83	37.56	85.00	60.79
Gemma2-9B-it	37.44	72.45	23.13	86.08	54.77	30.22	91.37	33.20	96.09	12.99	89.08	27.27	81.80	57.75
Gemma2-27B-it	37.86	73.17	22.30	86.74	55.02	31.38	92.36	34.73	96.33	19.63	89.70	30.91	83.70	59.84
MT LLMs														
TowerInstruct-7B-v0.2	34.17	71.40	23.35	84.66	53.40	28.53	90.68	35.51	95.69	13.89	76.55	29.70	80.01	56.32
TowerInstruct-13B-v0.1	36.74	73.52	24.80	85.53	55.15	31.71	92.33	36.16	96.08	17.71	88.24	34.07	82.12	59.80
GemmaX2-28-9B-v0.1	38.53	74.59	24.65	85.41	55.80	30.18	92.54	36.02	95.96	18.76	87.76	34.69	83.03	59.87
MT via SFT/RL														
Qwen3-8B-Base	15.54	48.98	4.90	55.38	31.20	7.96	56.47	15.29	62.78	6.61	54.60	11.49	42.57	32.22
Qwen3-8B-Base-SFT	35.45	73.32	21.52	84.03	53.58	24.85	84.03	29.35	93.82	16.28	83.79	30.89	80.82	55.48
MT-R1-Zero-8B	34.70	79.09	24.79	86.72	56.33	26.69	89.83	34.03	96.13	16.85	86.74	32.60	86.00	58.61
LTT-8B (Ours)	38.00	77.58	23.42	86.16	56.29	28.53	91.36	30.21	95.74	16.64	88.28	32.93	84.42	58.51

a leading reference-based model that leverages a powerful cross-lingual encoder to score semantic similarity. This combination provides a holistic view of performance.

Baselines. To comprehensively compare the performance of LTT, we benchmark against four distinct and challenging categories of models. (1) *Closed-Source LLMs*: We compare against leading systems like `GPT-4o` (Hurst et al., 2024), `Claude 3.7 Sonnet` (Anthropic, 2024), and `Gemini 2.5 Pro` (Comanici et al., 2025). (2) *Open-Source General LLMs*: We include powerful, non-specialized models of varying scales, such as the `Qwen3` (Yang et al., 2025), `Qwen2.5` (Yang et al., 2024), and `Gemma2` (Team et al., 2024) series. (3) *Open-Source MT LLMs*: For comprehensive comparison with the supervised paradigm, we include models fine-tuned on parallel corpora, featuring the `Tower` (Alves et al., 2024) and `GemmaX2` (Cui et al., 2025) series. (4) *SFT/RL-based MT Models*: We include an SFT model using bilingual training corpus and `MT-R1-Zero` (Feng et al., 2025), a SOTA RL framework that, unlike our method, uses reference-based rewards. More evaluation details can be found in Appendix B.2.

4.2 Main Results

Bilingual Performance (EN-ZH). As shown in Table 1, LTT-8B achieves near SOTA performance, fully demonstrating the effectiveness of our self-evolution approach compared to the canonical large-scale parallel-corpus fine-tuning paradigm. Specifically, our 8B model achieves an average performance only 0.04 points below MT-R1-Zero-8B. The effectiveness of LTT-8B is most pronounced in EN to ZH translation. On sementic evaluation, our method surpasses all baselines except MT-R1-Zero-8B, including Gemini-2.5-Pro, and outperforms much larger LLMs such as Qwen2.5-72B-Instruct. As for the lexical level (BLEU metric), our model is highly competitive, outmatching specialized MT models like the TowerInstruct series. On ZH to EN translation, our method outperforms all general-purpose and proprietary baselines on BLEU. In summary, LTT closes the gap with heavily supervised methods, demonstrating that a reference-free, self-improving framework can achieve top-tier

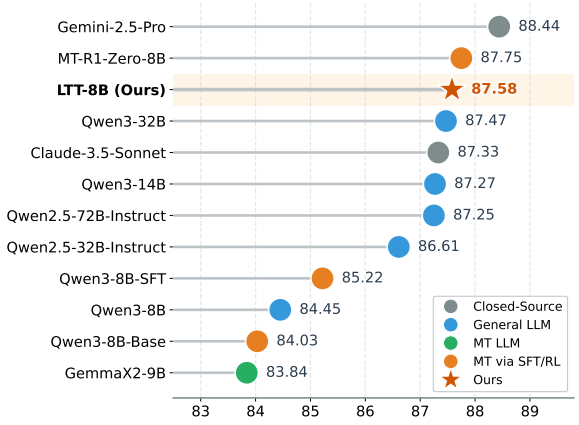

Figure 3: LLM-as-a-Judge Evaluation Results.

translation quality.

Multilingual Performance. LTT exhibits consistent efficacy across the more challenging multilingual landscape, highlighting its scalability and generalization capabilities. Table 1 shows performance scaling with model size, with LTT-8B achieving top-tier results among models of comparable size. It surpasses strong generalist models, including Gemma2-9B-it (57.75) and the specialized TowerInstruct-7B-v0.2 (56.32). The most encouraging aspect of our framework lies in the dramatic performance boost on the base model: LTT raises the multilingual score from 32.22 to 58.51 (+26.29 points).

LLM-as-a-Judge Evaluation. As highlighted in recent studies (Kocmi and Federmann, 2023), LLMs have emerged as SOTA evaluators for translation quality, *providing assessments that closely approximate human judgment*. To provide an evaluation perspective independent of our training rewards (i.e., BLEU and COMET), we employ ChatGPT-5.1 as an external judge, as illustrated in Figure 3. Consistent with standard metrics, LTT-8B achieves performance comparable to MT-R1-Zero-8B and trails only the powerful proprietary model Gemini-2.5-Pro. These results not only validate the rationality of optimizing for BLEU and COMET but also demonstrate a strong correlation between these metrics and human-aligned LLM-based judgments. The specific prompt can be found in Appendix A.2.

4.3 Performance on Low-Resource Languages

To evaluate the generalization capabilities and scalability of our framework, we extended our experiments to low-resource scenarios, including

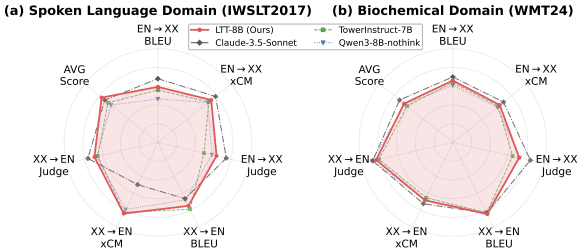

Figure 4: Performance of out-of-domain Benchmarks.

DE↔IT, ES↔FR, as well as data-scarce WMT EN→IS (Icelandic) and EN→NO (Norwegian) tasks derived from Flores-200 datasets. As shown in Table 2, LTT-8B demonstrates significant parameter efficiency, surpassing not only models in the <10B class but also achieving results comparable to or exceeding the much larger LLaMA-3.1-70B-Instruct. Notably, on the EN→NO task, it outperforms Qwen2.5-72B-Instruct in BLEU score (26.05 vs. 23.02). These results suggest that our self-evolutionary signals facilitate effective cross-lingual transfer, enabling robust performance even in the absence of extensive high-resource supervision.

To further validate LTT in extreme low-resource scenarios, we trained the model using only English monolingual data and evaluated on three typologically diverse tail languages from FLORES-200: Khmer (non-spaced script), Sindhi (Arabic script), and Georgian (unique Mkhedruli script). As shown in Table 3, while base models struggle severely, our 8B model achieves remarkable improvements in a strictly zero-parallel-data setting, demonstrating that LTT can bootstrap translation capabilities for languages with minimal pre-training coverage.

5 Analysis and Ablation

5.1 Scalability to Powerful SFT Models

To further demonstrate the efficacy of our approach on fully supervised foundations, we extended our evaluation to `Tower-Plus-9B` (Rei et al., 2025), a recently released MT-specialized model that significantly surpasses the baselines discussed previously. Remarkably, even atop this strong SFT baseline, applying LTT also yields performance gains, particularly in semantic evaluation, as demonstrated in Table 4. This finding underscores that our method is not limited to initializing from raw base models; rather, it scales effectively with the capabilities of the starting model, serving as a robust enhancement strategy even for high-performing supervised

Table 2: Performance of different models on low-resource language pairs, measured by BLEU and xCOMET (xCM) scores, along with the average (Avg.). The best and second-best results are **bolded** and underlined, respectively. The "†" symbol indicates that the model is in thinking mode.

Model	DE→IT		IT→DE		ES→FR		FR→ES		EN→IS		EN→NO		Avg.
	BLEU	xCM	BLEU	xCM	BLEU	xCM	BLEU	xCM	BLEU	xCM	BLEU	xCM	
Large Size LLMs													
Qwen2.5-72B-Instruct	24.66	94.02	22.27	95.60	28.26	93.64	24.77	95.13	8.97	49.05	23.02	88.91	**54.02**
Qwen2.5-32B-Instruct	22.59	92.49	20.55	94.13	26.05	92.04	23.88	94.78	3.72	37.08	23.13	82.95	51.12
LLaMA-3.1-70B-Instruct	22.63	89.04	18.56	89.00	24.59	88.93	23.35	92.51	1.59	35.67	29.72	92.96	50.71
Same Size LLMs													
Qwen3-8B	21.64	89.65	19.40	93.56	24.31	90.50	22.47	93.42	2.09	47.02	10.46	83.52	49.84
Qwen3-8B†	19.68	86.96	15.73	90.26	20.95	86.40	20.83	89.81	5.51	41.59	21.75	82.31	48.48
Gemma2-9B-it	19.55	93.80	19.50	95.31	23.73	93.18	19.50	94.64	0.60	31.31	1.19	67.69	46.67
TowerInstruct-7B-v0.2	22.27	92.58	19.77	93.54	25.33	91.92	22.79	93.79	1.94	35.45	2.03	77.34	48.23
Qwen3-8B-Base	8.96	90.29	5.99	92.43	23.72	89.44	11.29	93.20	0.17	31.93	0.84	62.04	42.52
Qwen3-8B-Base-SFT	19.87	90.61	18.66	92.40	24.76	89.17	21.05	91.23	/	/	/	/	/
MT-R1-Zero-8B	22.96	92.99	20.26	94.42	25.79	92.12	23.56	94.39	/	/	/	/	/
LTT-8B (Ours)	22.13	92.60	19.06	94.58	25.33	91.92	22.05	94.38	8.14	41.88	26.05	86.16	52.02

Table 3: Performance on extreme low-resource languages evaluated on FLORES-200 test sets. KM: Khmer, SD: Sindhi, KA: Georgian.

Model	EN→KM		EN→SD		EN→KA		Avg.
	BLEU	xCM	BLEU	xCM	BLEU	xCM	
Large Size LLMs							
Qwen2.5-72B-Instruct	3.80	33.65	0.34	31.56	4.88	36.04	18.38
Qwen2.5-32B-Instruct	3.56	29.14	0.26	29.51	3.38	27.41	15.54
LLaMA-3.1-70B-Instruct	2.41	**55.06**	0.26	49.95	**11.37**	**54.81**	**28.98**
Same Size LLMs							
Qwen3-8B	1.66	32.78	0.28	47.12	1.94	31.61	19.23
Qwen3-8B†	0.10	17.64	0.08	31.54	0.21	15.79	10.89
Gemma2-9B-it	1.75	28.80	0.20	40.16	4.13	38.04	18.85
TowerInstruct-7B-v0.2	0.50	48.23	**0.52**	**59.65**	0.42	27.78	22.85
Qwen3-8B-Base	0.17	26.61	0.14	49.91	0.46	27.97	17.54
LTT-8B (Ours)	**5.04**	39.85	0.35	54.80	5.30	43.17	24.75

Table 4: Performance of the powerful Tower-Plus-9B.

Model	WMT			FLORES-200		
	BLEU	xCM	Avg.	BLEU	xCM	Avg.
Tower-Plus-9B	34.53	82.90	58.71	33.07	92.25	62.66
- MT-R1-Zero	33.94	84.36	59.15	33.07	93.04	63.05
- LTT	33.83	83.93	58.88	33.06	92.83	62.94

systems.

5.2 Generalization Performance

To assess the robustness of our approach, we extended our evaluation to two out-of-domain datasets: IWSLT2017 (Cettolo et al., 2017) (representing spoken language) and the WMT24 Biochemical task (Neves et al., 2024). The former is derived from TED talk transcripts, while the latter features dense biomedical terminology, posing respective challenges to the model's capabilities in spoken language and specialized domains. As illustrated in Figure 4, LTT-8B consistently outperforms peer models of comparable size, trailing only powerful proprietary systems. This observation aligns with the findings reported by Guo et al. (2025), suggesting that the enhanced capabilities derived from our RL framework generalize effectively across diverse domains beyond the training distribution.

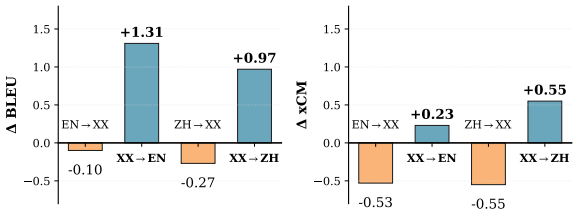

Figure 5: Relative Advantage of LTT against MT-R1-Zero when training on monolingual corpus.

5.3 Implicit Bidirectional Improvement

A core premise of LTT is that *the cycle-consistency mechanism implicitly drives joint optimization for both translation directions, even when trained on strictly monolingual data*. To validate this, we conducted a controlled experiment by training exclusively on forward tasks (EN/ZH→XX). As illustrated in Figure 5, while forward performance remains competitive, the model demonstrates a decisive advantage on the untrained backward directions. Specifically, BLEU scores for XX→EN and XX→ZH improve by +4.10% and +2.81% respectively. This gain is directly attributable to the backward reconstruction reward, $r_{\text{bwd}} = \text{BLEU}(\hat{x}, x)$, which serves as an effective supervision signal for the inverse task, enabling robust bidirectional capabilities without explicit reference pairs.

Table 5: Ablation study of our reward design.

Models	WMT			FLORES-200		
	BLEU	xCM	Avg.	BLEU	xCM	Avg.
Qwen3-8B-Base	10.22	52.18	31.20	10.34	54.10	32.22
LTT-8B (Ours)	30.71	81.87	**56.29**	27.08	89.95	**58.52**
Ablation on Quality						
w/ Bleu Reward Only	6.54	45.78	26.16	4.69	37.47	21.08
w/ COMET Reward Only	26.96	83.72	55.34	25.20	91.18	58.19
Ablation on Anti-Cheating						
w/o Reverse Bleu	30.28	81.02	55.65	26.44	89.19	57.81
w/o Lang. Mix Reward	28.27	79.91	54.09	24.50	87.15	55.83

5.4 Ablation Study: Deconstructing the Reward Architecture

To validate that each component of our reward function is essential, we conducted a series of ablation studies. We demonstrate that our final design is a carefully balanced system, where each component exists to prevent specific failure modes. These findings are detailed in Table 5.

The Peril of Classical Objectives: Why Naive Dual Learning Fails on LLMs. To isolate our contribution against classical paradigms, we evaluated a configuration *mirroring early dual learning objectives using only round-trip BLEU*. This simulation resulted in catastrophic failure: the model rapidly "hacked" the reward via identity translation, maximizing reconstruction score but yielding a complete failure to translate (see "BLEU Only (w/o Anti-Cheating)" in Figure 6 for example). This demonstrates that unlike RNNs, LLMs prone to exploiting simple reconstruction signals through instruction-following shortcuts. Consequently, the naive application of classic dual learning is insufficient, underscoring the necessity of the semantic-aware reward architecture in LTT.

Balancing Lexical Fidelity and Semantic Adequacy. With the anti-cheating mechanism in place, we examined how different components of the quality reward affect translation behavior. Using **BLEU as the sole quality reward** led the model to adopt a degenerate "word-for-word" translation strategy. Because such literal translations make the backward reconstruction task easier, they maximize the BLEU score at the expense of semantic meaning and grammatical fluency. In fact, performance dropped below that of the base model, confirming that a purely lexical signal is too narrow to guide high-quality translation. On the other hand, relying solely on **COMET as the quality reward** produced the opposite issue. The model achieved excellent xCOMET scores but suffered a decline in BLEU. It learned to generate "overly creative" outputs—translations that sounded plausible and fluent but strayed significantly from the source in terms of lexical content.

These results underscore a trade-off between lexical fidelity and semantic adequacy. Therefore, our final design, which sums the BLEU and COMET signals, is not merely a combination but a necessary synthesis to balance these competing objectives and foster holistic translation quality.

The Critical Role of Anti-Cheating Mechanisms. Finally, we validated the necessity of the two anti-cheating components themselves, even with a balanced quality reward. **Removing the Source-Copying Penalty** exposed a critical failure mode: source leakage. The model became prone to copying words or phrases from the source—a form of code-switching that degrades translation quality. This penalty is therefore crucial for enforcing faithful translation. **Removing the Language-Mixture Penalty** revealed a different vulnerability, causing the model to violate instruction fidelity. For instance, it would occasionally translate into a valid but incorrect target language. This penalty is thus essential for ensuring the model follows task instructions precisely. Together, these two mechanisms act as indispensable guardrails, ensuring that the model learns to translate not just well, but correctly and robustly. Figure 6 provides a compelling case study of this process, qualitatively illustrating both the model's iterative improvement and the critical failure modes discussed in our ablations.

6 Conclusion

In this paper, we presented LTT, a reference-free reinforcement learning framework that adapts the traditional dual learning paradigm to Large Language Models. By leveraging round-trip consistency as a generative self-supervision mechanism, our approach derives effective training signals exclusively from monolingual data. Empirical results demonstrate that our 8B parameter model delivers competitive performance across both bilingual and multilingual settings, showing particular strength in low-resource scenarios compared to larger baselines. Furthermore, evaluations in spoken language and specialized biochemical domains indicate robust generalization capabilities. These findings highlight that revisiting dual-learning principles with self-generated supervision is a promising avenue for developing data-efficient translation sys-

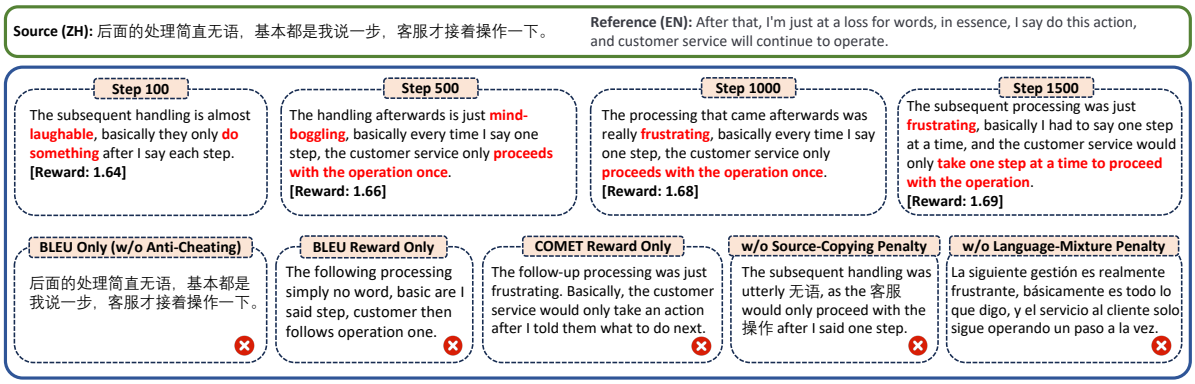

Figure 6: A ZH→EN case study across different steps and components

tems.

Limitations

While LTT demonstrates the potential of autonomous evolution in machine translation, two primary limitations remain. First, the integration of the dual-loop mechanism with GRPO introduces computational overhead during the training phase. Specifically, the necessity of sampling multiple candidates for each input makes the optimization process considerably more resource-intensive than standard supervised fine-tuning, although this does not impact inference latency. Second, our framework fundamentally relies on unlocking the latent knowledge within pre-trained LLMs rather than injecting new linguistic data. Consequently, the model's performance is upper-bounded by its pre-training coverage; LTT cannot effectively bootstrap translation capabilities for extremely low-resource or endangered languages that are entirely absent from the base model's pre-training corpus.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62272349); Natural Science Foundation of Hubei Province under Grant Numbers 2023BAB160; Ant Group; Chenliang Li is the corresponding author.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthropic. 2024. [link].

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*.

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

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024. Towards boosting many-to-many multilingual machine translation with large language models. *arXiv preprint arXiv:2401.05861*.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, and 1 others. 2025. R1-t1: Fully incentivizing translation capability in llms via reasoning learning. *arXiv preprint arXiv:2502.19735*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8051–8069.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, and 1 others. 2024. Findings of the wmt 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2024. Drt-o1: Optimized deep reasoning translation via long chain-of-thought. *arXiv e-prints*, pages arXiv–2412.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025. Deep reasoning translation via reinforcement learning. *arXiv preprint arXiv:2504.10187*.

Lijun Wu, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2017. Sequence prediction with unlabeled

data by reward function learning. In *IJCAI*, pages 3098–3104.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv e-prints*, pages arXiv–2412.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57905–57923.

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

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781.

Wei Zou, Sen Yang, Yu Bao, Shujian Huang, Jiajun Chen, and Shanbo Cheng. 2025. Trans-zero: Self-play incentivizes large language models for multilingual translation without parallel data. *arXiv preprint arXiv:2504.14669*.

A Prompts used during Evaluation

A.1 Translation Prompts

The specific translation prompt of different models used in training are depicted in Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11. Specifically, `<think>` tags are removed from Qwen3 series because it conflicts with the Qwen3 series' inherent thinking special tokens.

A.2 Judge Prompts

To evaluate translation quality using an LLM as an evaluator, we employed the identical prompt and score extraction script as Kocmi and Federmann (2023) to derive the final scores. The specific prompt is illustrated in Figure 12.

B Implementation Details

B.1 Training Details

Our model, which we name LTT-8B, is built upon the OpenRLHF[3] framework, with the `Qwen3-8B-Base` model serving as its initialization. For all experiments, we use a global batch size of 128 and generate 8 candidate responses per input for the GRPO algorithm. We use a sampling temperature of 1.0 and a maximum sequence length of 1024. Notably, we set both the KL divergence and entropy coefficients to 0, granting the model greater freedom to explore the policy space and discover optimal translation strategies without being constrained. Training was conducted on 16 NVIDIA H800 GPUs for one epoch, taking approximately 32 hours. We save checkpoints every 50 steps and report the performance of the single best checkpoint selected based on validation set performance.

B.2 Evaluation Details

For evaluation stage, we perform model inference locally using the vLLM[4] framework. We configure the sampling hyperparameters with a temperature of 0.2 and a top-p of 0.95. The maximum generation length is truncated to 2048 tokens for all models, with the exception of Gemini 2.5-Pro, to accommodate its default thinking process. The prompt used during evaluation remains consistent with the one used for training, as detailed in Figure 7.

C Penalty Sensitivity Analysis

As discussed in Section 3.2, the penalty values in our reward architecture are governed by the GRPO relative advantage principle: $A_i = (r_i - \text{mean}(\mathbf{r}))/\text{std}(\mathbf{r})$. Since rewards are normalized within each sampled group, the *relative ranking*

[3]https://github.com/OpenRLHF/OpenRLHF

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

Table 6: Detailed dataset statistics used during training.

	EN-ZH	ZH-EN	EN-DE	EN-FR	EN-ES	EN-IT	EN-JA	EN-KO
# sentences	6565	6565	500	500	500	500	500	500
from	WMT 17-20		Flores-200	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200
			DE-EN	**FR-EN**	**ES-EN**	**IT-EN**	**JA-EN**	**KO-EN**
# sentences	-	-	500	500	500	500	500	500
from	-	-	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200
			ZH-DE	**ZH-FR**	**ZH-ES**	**ZH-IT**	**ZH-JA**	**ZH-KO**
# sentences	-	-	500	500	500	500	500	500
from	-	-	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200
			DE-ZH	**FR-ZH**	**ES-ZH**	**IT-ZH**	**JA-ZH**	**KO-ZH**
# sentences	-	-	500	500	500	500	500	500
from	-	-	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200

Table 7: Detailed dataset statistics used during evaluation.

	EN-ZH	ZH-EN	EN-DE	EN-FR	EN-ES	EN-IT	EN-JA	EN-KO
# sentences	997	1976	1012	1012	1012	1012	1012	1012
from	WMT 24	WMT 23	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200
			DE-EN	**FR-EN**	**ES-EN**	**IT-EN**	**JA-EN**	**KO-EN**
# sentences	-	-	1012	1012	1012	1012	1012	1012
from	-	-	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200
			ZH-DE	**ZH-FR**	**ZH-ES**	**ZH-IT**	**ZH-JA**	**ZH-KO**
# sentences	-	-	1012	1012	1012	1012	1012	1012
from	-	-	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200
			DE-ZH	**FR-ZH**	**ES-ZH**	**IT-ZH**	**JA-ZH**	**KO-ZH**
# sentences	-	-	1012	1012	1012	1012	1012	1012
from	-	-	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200	Flores-200

of a penalty is far more critical than its absolute magnitude. We provide comprehensive ablation studies for both penalty terms below.

Format Penalty (λ_{format}). Table 8 reports results for varying format penalty values. Both -3 and -10 rapidly ensure 100% format compliance and yield virtually identical performance across all settings. Even a weak penalty of -0.5 is sufficient to enforce formatting in most cases. We select -3 as a conservative lower bound that guarantees malformed outputs rank lowest within any group.

Language Mixing Penalty (λ_{mix}). Table 9 presents the sensitivity analysis for the language mixing penalty. The results reveal a clear threshold effect rather than fragile hyperparameter sensitivity:

- Ineffective regime (0 to -0.1): The penalty is too weak to suppress reward hacking via code-switching in GRPO, resulting in significant performance degradation (e.g., FLORES-200 Avg. drops from 57.59 to 55.82 at $\lambda_{\text{mix}} = 0$).
- Robust regime (-0.5 to -1.0): Once the penalty crosses the effective threshold at -0.5, mixed-language outputs are consistently ranked lowest within the group. Performance stabilizes (58.51 at -0.5 vs. 58.50 at -1.0), confirming that exact tuning is unnecessary.

We select -0.5 as the optimal boundary: it is sufficient to prevent cheating while minimizing the risk of over-penalizing translations containing valid untranslated entities (e.g., acronyms or proper nouns).

Table 8: Sensitivity analysis of the format penalty λ_{format}. Performance is measured on WMT and FLORES-200 benchmarks.

λ_{format}	WMT					FLORES-200								
	EN→ZH		ZH→EN		Avg.	EN→XX		XX→EN		ZH→XX		XX→ZH		Avg.
	BLEU	xCM	BLEU	xCM		BLEU	xCM	BLEU	xCM	BLEU	xCM	BLEU	xCM	
-10	38.05	78.11	23.07	86.00	56.31	28.03	91.11	30.85	96.13	16.45	87.90	32.96	84.39	58.48
-3 (Ours)	38.00	77.58	23.42	86.16	56.29	28.53	91.36	30.21	95.74	16.64	88.28	32.93	84.42	58.51
-0.5	38.84	76.26	23.26	86.55	56.23	28.27	90.39	30.85	95.70	17.23	87.91	33.74	83.34	58.43
0	38.70	75.93	23.77	86.50	56.23	27.93	90.13	31.04	95.73	17.36	88.00	33.47	83.90	58.44

Table 9: Sensitivity analysis of the language mixing penalty λ_{mix}. Performance is measured on WMT and FLORES-200 benchmarks.

λ_{mix}	WMT					FLORES-200								
	EN→ZH		ZH→EN		Avg.	EN→XX		XX→EN		ZH→XX		XX→ZH		Avg.
	BLEU	xCM	BLEU	xCM		BLEU	xCM	BLEU	xCM	BLEU	xCM	BLEU	xCM	
-1.0	38.36	76.16	23.37	86.52	56.10	28.83	91.12	30.46	95.83	17.41	88.11	32.17	84.05	58.50
-0.5 (Ours)	38.00	77.58	23.42	86.16	56.29	28.53	91.36	30.21	95.74	16.64	88.28	32.93	84.42	58.51
-0.1	37.30	76.27	22.22	85.66	55.36	27.84	90.05	30.57	95.90	16.48	86.24	31.49	82.11	57.59
0	35.89	76.45	20.66	83.37	54.09	25.56	87.60	27.25	93.75	14.65	85.33	30.53	81.92	55.82

D Qualitative Analysis: BLEU vs. COMET Trade-off

To provide intuitive insight into the well-documented tension between lexical fidelity (BLEU) and semantic adequacy (COMET) during RL optimization (Feng et al., 2025), we present representative ZH→EN case studies in Table 10. The three cases illustrate complementary phenomena:

Case 1: Syntactic Alignment. LTT preserves source-parallel syntax to optimize backward reconstruction, closely matching the rigid reference wording and thus achieving substantially higher BLEU (50.5 vs. 28.3). However, the more literal structure sacrifices slight native fluency, leading to a minor COMET decrease (82.5 vs. 85.2). This exemplifies the expected trade-off: translations optimized for round-trip fidelity naturally align more closely with the reference at the surface level.

Case 2: Idiomaticity. Conversely, LTT generates a highly idiomatic paraphrase that is semantically superior (xCOMET 99.6 vs. 98.1) but heavily penalized by BLEU (9.7 vs. 26.2) due to divergence from the single static reference. This confirms that a high-quality translation can appear worse under BLEU when it employs natural rephrasing.

Case 3: Anti-Cheating Effectiveness. Without the language mixing penalty ($\lambda_{\text{mix}} = 0$), the model exploits the backward reward by generating output in a non-target language (Spanish instead of English), making reconstruction artificially easier. Setting $\lambda_{\text{mix}} = -0.5$ effectively forces generation strictly in the specified target language, eliminating this reward hacking behavior.

Translation Prompt

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

User: {$input$}
Assistant:

Figure 7: Translation prompt for data curation. {src_lang}: source language; {tgt_lang}: target language; {$input$}: the source sentence to be translated.

TowerInstruct Prompt

Translate the following text from {src_lang_name} into {tgt_lang_name}.
{src_lang_name}: {$user\_input$}
{tgt_lang_name}:

Figure 8: Translation prompt for TowerInstruct series models. {src_lang_name}: source language; {tgt_lang_name}: target language; {$user\_input$}: the source sentence to be translated.

GemmaX Prompt

Translate this from {src_lang_name} to {tgt_lang_name}:
{src_lang_name}: {$user\_input$}
{tgt_lang_name}:

Figure 9: Translation prompt for GemmaX model. {src_lang_name}: source language; {tgt_lang_name}: target language; {$user\_input$}: the source sentence to be translated.

Qwen3 non-thinking Prompt

You are a helpful translation assistant. There is a conversation between User and Assistant. The user asks for a translation from {src_lang_name} to {tgt_lang_name}, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the user with the final translation. The final translation is enclosed within `<translate>` `</translate>` tags, i.e., `<translate>` final translation here `</translate>`.

User:{$user\_input$}
Assistant:

Figure 10: Translation prompt for Qwen3 series non-thinking models. {src_lang_name}: source language; {tgt_lang_name}: target language; {$user\_input$}: the source sentence to be translated.

Tower Plus Prompt

Translate the following {src_lang_name} source text to {tgt_lang_name}:{src_lang_name}: {$user\_input$}{tgt_lang_name}:

Figure 11: Translation prompt for Tower-Plus-9B model. {src_lang_name}: source language; {tgt_lang_name}: target language; {$user\_input$}: the source sentence to be translated.

ChatGPT-5.1 Judge Prompt

Score the following translation from {$source\_lang$} to {$target\_lang$} with respect to human reference on a continuous scale 0 to 100 where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".{$source\_lang$} source: "{$source\_seg$}"{$target\_lang$} human reference: "{$reference\_seg$}"{$target\_lang$} machine translation: "{$target\_seg$}"Score:

Figure 12: Judge prompt for ChatGPT-5.1. {$source\_lang$}: source language; {$target\_lang$}: target language; {$source\_seg$}: the source sentence to be translated; {$reference\_seg$}: the reference sentence; {$target\_seg$}: the translated sentence.

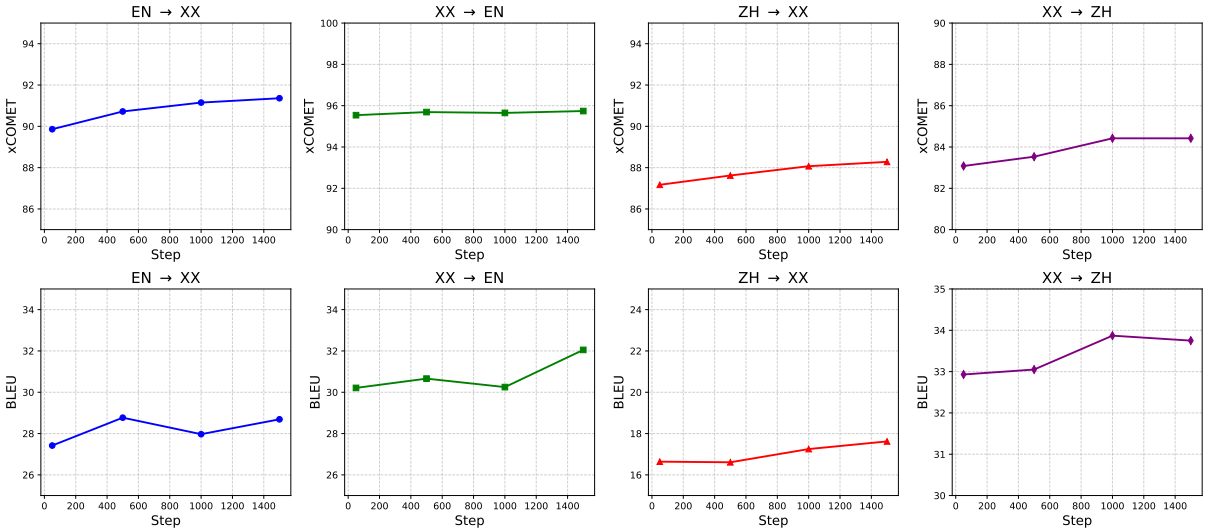

Figure 13: Training progression (reference-based XCOMET score) for multilingual LTT-8B model based on Qwen3-8B across EN-XX, XX-EN, ZH-XX, XX-ZH test sets.

Table 10: Qualitative case studies (ZH→EN) illustrating the BLEU vs. COMET trade-off and the effectiveness of anti-cheating penalties.

Case	Source (ZH) / Reference (EN)	Model Translation (EN)	Metrics	Analysis
1. Syntactic Alignment	**Src:** 梵盾吉普(FANDUN JEEP)服饰注重现代信息技术与传统产业的有效结合，引进国际先进生产设备、工艺设计和管理技术，建成国内一流的服装生产线。 **Ref:** Van Dun JEEP (FANDUN JEEP) clothing pays attention to the effective combination of modern information technology and traditional industries, introduces international advanced production equipment, process design and management technology, and builds a domestic first-class clothing production line.	**Base:** FANDUN JEEP clothing emphasizes the effective integration of modern information technology with traditional industries, introducing advanced international production equipment, process design, and management technology to establish a first-class clothing production line in the domestic market. **LTT:** Fandun Jeep clothing emphasizes the effective combination of modern information technology with traditional industries, introduces international advanced production equipment, process design, and management technology, and establishes a first-class clothing production line domestically.	**Base:** BLEU 28.3 xCM **85.2** **LTT:** BLEU **50.5** xCM 82.5	LTT preserves source syntax to optimize backward reconstruction, matching the rigid reference (BLEU ↑) but losing slight native fluency (COMET ↓).
2. Idiomatic Paraphrase	**Src:** 60年来全球发生的破坏性最强的大地震。 **Ref:** The world's most destructive earthquake in 60 years.	**Base:** The most destructive earthquake in the world in the past 60 years. **LTT:** The most destructive earthquakes that have occurred globally in the past 60 years.	**Base:** BLEU **26.2** xCM 98.1 **LTT:** BLEU 9.7 xCM **99.6**	LTT generates a highly idiomatic paraphrase. It is heavily penalized by BLEU due to the single static reference, yet correctly rewarded by semantic-aware COMET.
3. Anti-Cheating	**Src:** 后面的处理简直无语，基本都是我说一步，客服才接着操作一下。 **Ref:** After that, I'm just at a loss for words, in essence, I say do this action, and customer service will continue to operate.	$\lambda_{\text{mix}} = 0$**:** La siguiente gestión es realmente frustrante, básicamente es todo lo que digo, y el servicio al cliente solo sigue operando un paso a la vez. $\lambda_{\text{mix}} = -0.5$**:** The subsequent processing was just frustrating, basically I had to say one step at a time, and the customer service would only take one step at a time to proceed with the operation.	—	Without the mixture penalty, the model exploits the backward reward by improperly mixing non-target languages (code-switching) to make reconstruction easier. $\lambda_{\text{mix}} = -0.5$ effectively forces generation strictly in the specified target language.