Breaking Consensus Bias: Unsupervised Reinforcement Learning for Machine Translation

Shuting Jiang[1,2], Ran Song[1,2*], Siqi Zhang[1,2], Yuxin Huang[1,2], Shengxiang Gao[1,2], Zhengtao Yu[1,2*]
[1] Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China
[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
{shuting_jiang22, song_ransr, huangyuxin2004}@163.com,
zhangsiqi.mo@foxmail.com, {gaoshengxiang.yn,ztyu}@hotmail.com

Abstract

Reinforcement learning (RL) excels in reasoning tasks with verifiable rewards, while its adaptation to machine translation (MT) remains challenging due to the lack of unique reward signals under multiple valid translations. Existing RL approaches for MT face either fixed references in supervised settings or the production of homogeneous references leading to mode collapse in unsupervised settings. Both limitations arise from ignoring entropy dynamics in RL-based MT. The core challenge is leveraging entropy for supervision construction and self-evolution. In this paper, we propose an Entropy-Driven Unsupervised RL for MT. Our framework integrates entropy-guided sampling for exploration, confidence-weighted label generation to transcend majority-voting bias, and uncertainty-aware optimization to prioritize high-entropy tokens. These mechanisms allow reward signals to co-evolve with model proficiency beyond fixed references. Experiments across multiple language pairs show our method outperforms supervised and unsupervised baselines by +0.63 and +2.52 average points, respectively. Our code is available at https://github.com/fortunatekiss/URLMT.

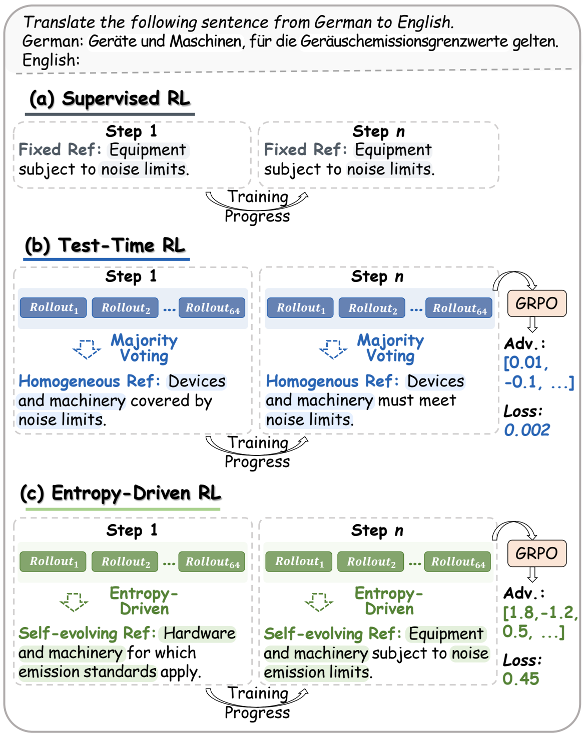


Figure 1: Comparison of RL Paradigms. (a) **Supervised RL** relies on fixed references. (b) **Test-Time RL** generates homogeneous references. (c) **Our Entropy-Driven RL** leverages entropy to generate diverse and self-evolving references.

1 Introduction

Reinforcement Learning (RL) has catalyzed significant breakthroughs in Large Language Models (LLMs), particularly within the Reinforcement Learning with Verifiable Rewards (RLVR) paradigm (Wang et al., 2024b; Luong et al., 2024; Lambert et al., 2025; Guo et al., 2025; Yang et al., 2025a). In reasoning tasks like mathematical problem solving (Shao et al., 2024; Xiong et al., 2025; Team et al., 2025) and code generation (Gehring et al., 2024; Wang et al., 2024a; Li et al., 2024), outputs admit explicit, binary verification (e.g., passing unit tests or matching unique answers). This determinism yields stable, precise reward signals, facilitating robust policy optimization.

In contrast to reasoning tasks, MT outputs operate within a continuous, high-dimensional semantic space where a single source sentence maps to a manifold of valid translations (Cheng and Vlachos, 2024a; Kayano and Sugawara, 2025; Jiang et al., 2026). Current RL approaches largely rely on reward signals derived from fixed annotated references via metrics like BLEU, COMET or LLM-as-Judge evaluation (Feng et al., 2025; Li et al., 2025a). Ideally, the supervision should evolve; however, static references act as a low-level ceiling (*akin to limiting a student with 5th-grade poten-*

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 20798–20812

tial to 3rd-grade textbooks). As illustrated in Figure 1(a), a supervised RL model may converge to a fixed word in reference sentence like *equipment*, ignoring the contextually richer and more specialized *Geräte und Maschinen (equipment and machinery)*. Consequently, the static supervision imposes an artificial performance ceiling, preventing the model from transcending the limitations of the annotated references to achieve superior translation quality.

To transcend the limitations of static supervision, recent research has pivoted towards unsupervised, self-generated paradigms such as Test-Time RL (TTRL) (Zuo et al., 2025), which employ majority voting to construct reward signals. However, applying vanilla TTRL to MT inadvertently precipitates mode collapse (*akin to a student who selectively addresses trivial questions to secure a high surface-level score (BLEU) while evading complex challenges*). As illustrated in Figure 1(b), majority voting inherently favors high-probability, low-variance tokens, yielding "safe" yet homogenized translations. Crucially, this homogeneity causes the advantages estimated by GRPO to become negligible and indistinguishable. We contend that these pseudo-labels reflect the model's *consensus bias* rather than optimal quality. Consequently, the absence of distinctive learning signals prevents effective optimization, creating a detrimental feedback loop where the sampling space rapidly contracts.

In fact, entropy serves as a faithful proxy for the model's translation proficiency, as the model exhibits uncertainty among semantic alternatives in high-entropy positions (Liu et al., 2022; Cheng and Vlachos, 2024b). Accordingly, we posit that high-entropy positions deserve greater attention *(akin to a student advances by tackling complex hurdles rather than merely reviewing acquired knowledge)*. In summary, the core challenge is how to leverage entropy to construct superior supervision that drives the model toward iterative semantic refinement and self-evolution during RL training. As illustrated in Figure 1(c), our approach enables reward signals to co-evolve with the model's proficiency, ensuring that the self-generated label remains both challenging and attainable.

In this paper, we propose Unsupervised RL for MT, an entropy-driven self-evolving RL framework that overcomes the limitations of static references and consensus bias in MT. Specifically, **Entropy-based Dynamic Temperature Sampling:** We adjust the sampling temperature dynamically using token-level entropy to balance exploration and stability during translation. **Confidence-Weighted Self-Reinforcing Label Generation:** We incorporate the voting distribution with model confidence to select informative pseudo-labels beyond majority frequency, enabling reward signals to co-evolve with the model's competence. **Entropy-Aware Policy Optimization:** During policy updates, we re-weight the loss by entropy, prioritizing the optimization of challenging, high-uncertainty tokens over mastered ones. Our approach allows LLMs to derive adaptive reward signals in unsupervised MT, mitigating consensus dominance and local optima. Our main contributions are as follows:

- We analyze the fundamental limitations of existing RL paradigms for MT, identifying a *static reward ceiling* in reference-based supervision and *consensus-driven mode collapse* in TTRL, both arising from ignoring entropy dynamics in RL-based MT.
- We propose an **Entropy-Driven Unsupervised RL framework** for MT that leverages entropy as a proxy for uncertainty to construct self-evolving reward signals, enabling supervision to co-evolve with the model's translation competence.
- Extensive experiments show our method achieves highest average scores in all models and language pairs, outperforming the supervised and unsupervised RL baselines by +0.63 and +2.52 points.

2 Preliminaries

2.1 MT with LLMs

We formulate MT as a conditional sequence generation task. Given a source sentence $X = (x_1, \ldots, x_m)$, an LLM functions as a probabilistic policy π_θ parameterized by θ. It autoregressively generates a target sequence $Y = (y_1, \ldots, y_n)$ by modeling the joint probability:

$$p_\theta(Y \mid X) = \prod_{t=1}^{n} \pi_\theta(y_t \mid y_{<t}, X), \qquad (1)$$

where y_t denotes the token sampled from the vocabulary $\mathcal{V}$ at step t.

2.2 RL for LLM-based MT

Beyond supervised fine-tuning on parallel corpora, RL is employed to align the model with sequence-level quality metrics for better translation (He et al., 2025; Feng et al., 2025; Li et al., 2025a; Yang et al., 2025b; Song et al., 2025). After generating a complete translation Y, the model receives a

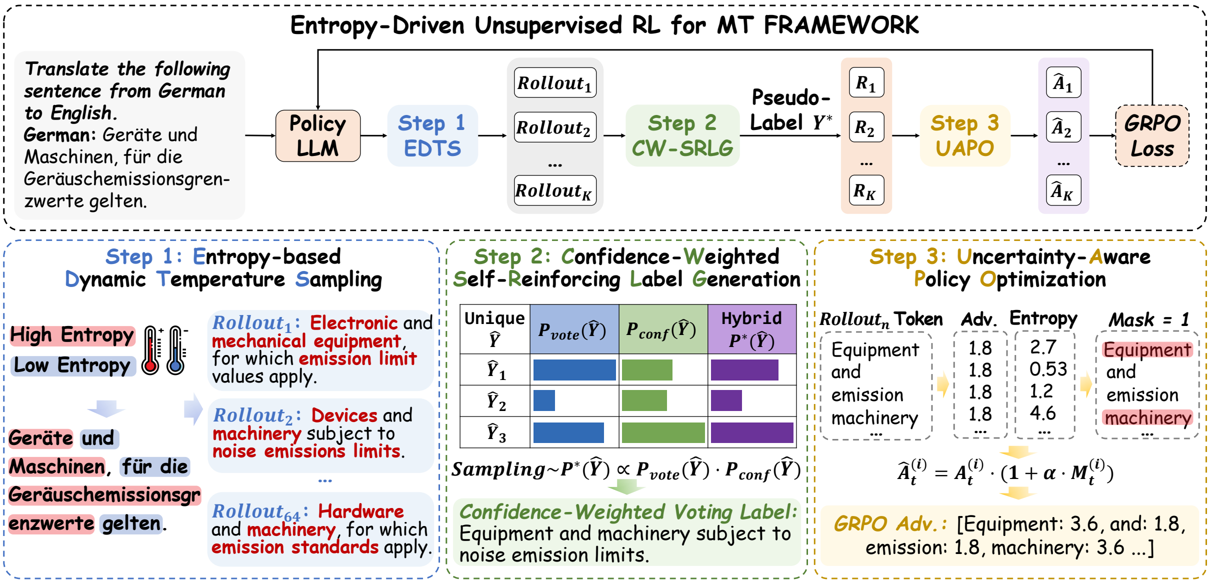


Figure 2: Overview of the Entropy-Driven Unsupervised RL framework. **EDTS** modulates sampling temperature based on token-level uncertainty, increasing it for high-entropy tokens and decreasing it for low-entropy ones. **CW-SRLG** generates self-evolving labels by weighting majority voting with model confidence. **UAPO** concentrates optimization on high-uncertainty positions to improve learning stability.

reward $R(Y, Y^*)$, typically derived from reference-based metrics (e.g., BLEU, COMET or the LLM-as-Judge evaluation) relative to reference Y^*. The objective is to maximize the expected reward:

$$\mathcal{J}(\theta) = \mathbb{E}_{Y \sim \pi_\theta(\cdot|X)}[R(Y, Y^*)]. \tag{2}$$

We follow the work of Feng et al. (2025) and Yang et al. (2025b) by adopting Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025) algorithm for training. For each input X, GRPO samples a group of K outputs $\{Y^{(i)}\}_{i=1}^K$ from the old policy $\pi_{\theta_{old}}$. The advantage $A^{(i)}$ for each output is computed by normalizing the rewards within the group:

$$A^{(i)} = \frac{R^{(i)} - \mu(\{R^{(j)}\}_{j=1}^K)}{\sigma(\{R^{(j)}\}_{j=1}^K)}, \tag{3}$$

where μ and σ denote the mean and standard deviation of the group rewards. The policy π_θ is updated by minimizing the following loss:

$$\mathcal{J}^{GRPO}(\theta) = \mathbb{E}_{X \sim P(X),\, \{Y^{(i)}\}_{i=1}^K \sim \pi_{\theta_{old}}(Y|X)} \Bigg[\frac{1}{K}\sum_{i=1}^K \min\Big(\frac{\pi_\theta(Y^{(i)}|X)}{\pi_{\theta_{old}}(Y^{(i)}|X)} A^{(i)}, \operatorname{clip}\Big(\frac{\pi_\theta(Y^{(i)}|X)}{\pi_{\theta_{old}}(Y^{(i)}|X)}, 1-\varepsilon, 1+\varepsilon\Big) A^{(i)}\Big) - \beta D_{KL}(\pi_\theta \| \pi_{ref}) \Bigg], \tag{4}$$

where ε and β are hyperparameters controlling the clipping threshold and the KL coefficient, respectively, and $D_{KL}(\pi_\theta \| \pi_{ref})$ is the KL divergence between π_θ and the reference model π_{ref}.

2.3 Test-Time RL

Standard RL requires ground-truth references for reward estimation, a prerequisite that becomes challenging when not having access to explicit label information. TTRL (Zuo et al., 2025) addresses this by synthesizing pseudo-labels during inference. The typical workflow involves: **(1) Sampling**: Sample K translation candidates $\{Y^{(i)}\}_{i=1}^K$ from $\pi_{\theta_{old}}$. **(2) Voting**: Determine a consensus output Y^{vote} from these candidates as the pseudo-label through majority voting. **(3) Optimization**: Reward $R(Y^{(i)})$ is defined as the similarity between $Y^{(i)}$ and Y^{vote}, i.e., $R(Y^{(i)}) = \text{Sim}(Y^{(i)}, Y^{\text{vote}})$, followed by policy update via Eq. 4.

3 Methodology

In this section, we detail the implementation of each stage for the entropy-driven unsupervised RL framework for MT. As illustrated in Figure 2, our framework consists of three stages: **(i) Entropy-based Dynamic Temperature Sampling** (EDTS), which modulates exploration breadth based on token-level uncertainty; **(ii) Confidence-Weighted Self-Reinforcing Label Generation** (CW-SRLG), which integrates majority voting consensus with model confidence to generate reliable supervision; and **(iii) Uncertainty-Aware Policy Optimization** (UAPO), which emphasizes learning from high-uncertainty tokens in MT.

3.1 Entropy-based Dynamic Temperature Sampling

To generate a diverse set of candidate samples, we enhance translation richness at the token level. In autoregressive generation, the sampling temperature T governs output diversity and is applied uniformly across tokens. However, uncertainty in MT is highly uneven: tokens with high translational complexity benefit from higher T for broader semantic exploration, whereas more deterministic tokens require lower T to maintain stability.

Since token-level output entropy serves as a reliable proxy for the model's internal uncertainty, we leverage it as a feedback signal to adaptively modulate the T at each decoding step. We calculate the adaptive temperature T_t as:

$$T_t = T_{\text{base}} \cdot (1 + \lambda \cdot \hat{z}_t), \tag{5}$$

where T_{base} is the baseline temperature, λ is a scaling factor, and $\hat{z}_t$ is the normalized uncertainty at t-th token. The uncertainty signal $\hat{z}_t$ is derived from the Shannon entropy H_t of the policy distribution over $\mathcal{V}$, defined as:

$$H_t = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid y_{<t}, X) \log \pi_\theta(v \mid y_{<t}, X). \tag{6}$$

In practice, entropy values follow a long-tailed distribution where a few difficult tokens exhibit unevenly high values while most deterministic tokens are concentrated near zero. To stabilize this, we first apply a log-transform $u_t = \log(H_t + \epsilon)$ to compress the range and then normalize this uncertainty metric as:

$$z_t = \frac{u_t - \tilde{\mu}}{\sigma + \epsilon}, \tag{7}$$

where $\tilde{\mu}$ and σ are the median and standard deviation of log-entropy values estimated from the training set. We clip z_t to $[-1, 1]$ to obtain $\hat{z}_t$. Accordingly, $\hat{z}_t > 0$ leads to an increase in T_t, encouraging semantic exploration at uncertain decoding steps, whereas $\hat{z}_t < 0$ reduces T_t to promote stable and confident predictions. Finally, each token is sampled using this dynamic T_t as:

$$y_t \sim \pi_\theta(\cdot \mid y_{<t}, X; T_t). \tag{8}$$

We perform K independent sampling runs for each X to obtain a candidate set $\mathcal{S} = \{Y^{(i)}\}_{i=1}^{K}$.

3.2 Confidence-Weighted Self-Reinforcing Label Generation

We next aim to select reliable pseudo-labels as supervision signals from the diverse translation candidates set $\mathcal{S}$. In MT, directly applying vanilla majority voting to generate pseudo-labels (as discussed in Sec. 2.3) tends to filter out high-confidence but infrequent candidates. Training with these consensus bias labels negates the value of exploration, leading to mode collapse and homogeneous outputs. To address this, we propose confidence-weighted voting, a mechanism that integrates majority voting consensus with intrinsic model confidence to construct high-quality pseudo-labels.

We define the confidence-weighted hybrid sampling distribution as:

$$P^*(\hat{Y}) = \frac{P_{\text{vote}}(\hat{Y}) \cdot P_{\text{conf}}(\hat{Y})}{Z}, \tag{9}$$

where $\hat{Y}$ denotes the unique candidate translation from $\mathcal{S}$, $P_{\text{vote}}(\hat{Y})$ is majority voting consensus distribution, $P_{\text{conf}}(\hat{Y})$ is intrinsic model confidence distribution, and Z is a normalization constant.

We first quantify $P_{\text{vote}}(\hat{Y})$ using standard majority voting based on its frequency in $\mathcal{S}$. For each $\hat{Y} \in \mathcal{S}$, the voting probability is defined as:

$$P_{\text{vote}}(\hat{Y}) = \frac{N(\hat{Y})}{K}, \tag{10}$$

where $N(\hat{Y})$ is the number of times $\hat{Y}$ appears in $\mathcal{S}$, and K is the total number of samples.

To counterbalance the consensus bias in $P_{\text{vote}}(\hat{Y})$, we incorporate the model's intrinsic confidence to measure the reliability of each $\hat{Y}$. Specifically, we derive the confidence distribution $P_{\text{conf}}(\hat{Y})$ by applying Softmax normalization to the length-averaged log-likelihoods:

$$P_{\text{conf}}(\hat{Y}) = \text{Softmax}\left(\frac{1}{|\hat{Y}|}\sum_{t=1}^{|\hat{Y}|} \log \pi_\theta(\hat{y}_t \mid \hat{y}_{<t}, X)\right). \tag{11}$$

This formulation prevents domination of P_{conf} by high-likelihood but low-information candidates.

Finally, we determine the pseudo-label Y^* by selecting the candidate that maximizes the confidence-weighted hybrid sampling distribution:

$$Y^* = \text{argmax}_{\hat{Y} \in \text{unique}(\mathcal{S})} P^*(\hat{Y}). \tag{12}$$

This sampling strategy preserves high-quality but lower-frequency candidates, effectively breaking the consensus bias loop inherent in vanilla TTRL.

Type	Method	ID Language Pairs								OOD Language Pairs											
		ZH⇒EN				EN⇒ZH				EN⇒JA				DE⇒EN				DE⇒ZH			
		Bl.	Co.	Kw.	Av.	Bl.	Co.	Kw.	Av.	Bl.	Co.	Kw.	Av.	Bl.	Co.	Kw.	Av.	Bl.	Co.	Kw.	Av.
Qwen2.5-0.5B-Instruct																					
-	Base	13.54	74.67	58.82	49.01	24.61	77.03	56.39	52.67	4.27	64.64	39.88	36.26	22.62	73.26	49.60	48.49	14.52	73.95	43.54	44.00
Sup.	GRPO	**16.21**	75.49	61.37	51.02	27.63	76.58	56.09	**53.68**	5.89	65.22	41.84	37.65	25.38	73.88	52.08	50.45	**19.36**	73.84	41.48	44.89
Sup.	DAPO	16.13	75.48	61.22	50.94	**27.64**	76.23	55.63	53.16	5.86	65.05	41.47	37.46	25.56	**74.11**	52.49	50.72	19.15	74.57	42.83	45.52
Unsup.	TTRL	13.25	73.96	57.13	48.11	26.11	76.02	55.85	52.66	4.82	63.65	38.55	35.67	22.28	73.28	49.24	48.26	16.67	73.10	41.92	43.89
Unsup.	**Ours**	16.07	**76.07**	**62.19**	**51.44**	27.15	**77.34**	**56.55**	**53.68**	**5.99**	**65.41**	**41.97**	**37.79**	**25.82**	73.84	**52.72**	**50.79**	19.08	**75.08**	**43.93**	**46.03**
Qwen3-4B-Instruct																					
-	Base	20.40	77.21	68.12	55.24	37.08	82.31	66.17	61.85	20.82	85.24	71.17	59.07	38.67	84.64	67.49	63.60	34.72	86.71	68.48	63.30
Sup.	GRPO	**23.89**	80.33	69.64	57.95	**40.76**	84.04	67.56	64.12	21.62	84.33	70.12	58.69	**41.71**	84.36	67.53	64.53	**36.38**	86.33	66.54	63.08
Sup.	DAPO	22.95	79.65	67.86	56.82	40.43	83.87	67.22	63.84	21.05	84.39	70.12	58.52	39.90	84.08	67.35	63.77	35.45	86.40	66.77	62.87
Unsup.	TTRL	19.18	79.19	68.50	55.62	37.16	82.53	66.04	61.91	20.12	83.69	69.19	57.67	39.57	84.42	67.55	63.84	34.89	85.82	67.35	62.68
Unsup.	**Ours**	23.37	**80.64**	**70.25**	**58.09**	39.97	**84.99**	**69.55**	**64.84**	**21.79**	**86.08**	**71.21**	**59.69**	41.09	**84.88**	**67.80**	**64.59**	35.83	**87.69**	**69.26**	**64.26**
Tower-Plus-2B																					
-	Base	24.41	80.54	70.04	58.33	37.54	82.69	62.63	59.25	22.95	80.20	68.56	57.23	42.76	78.89	65.69	62.44	35.75	79.18	60.23	58.38
Sup.	GRPO	24.77	80.43	70.07	58.42	**38.16**	82.71	67.39	62.75	23.07	83.47	71.14	59.22	43.38	80.66	67.25	63.76	35.40	83.21	64.07	60.89
Sup.	DAPO	**24.86**	80.45	69.99	58.43	37.29	82.78	67.32	62.46	23.10	83.49	71.02	59.20	**43.41**	81.00	**67.34**	63.91	35.66	83.44	64.03	61.04
Unsup.	TTRL	23.43	79.28	67.38	56.69	35.32	80.41	63.61	59.78	20.49	77.96	67.06	55.17	43.11	79.33	65.26	62.56	33.59	79.15	58.40	57.05
Unsup.	**Ours**	24.76	**82.83**	**72.63**	**60.07**	37.84	**83.62**	**67.58**	**63.01**	**23.11**	**83.62**	**71.34**	**59.36**	43.25	**81.41**	67.27	**63.97**	**35.82**	**84.28**	**65.15**	**61.75**

Table 1: Translation performance comparison using BLEU (Bl.), XCOMET (Co.) and COMETKiwi (Kw.) metrics, with average metric scores (Av.). Supervised (Sup.) and Unsupervised (Unsup.) labels indicate the learning setting. The **best** results are bolded and the second best results are underlined.

3.3 Uncertainty-Aware Policy Optimization

To better leverage uncertainty during policy updating, we propose an Uncertainty-Aware Policy Optimization method based on the obtained pseudo-label Y^*. In vanilla RL, every generated token receives the same scalar reward calculated by the entire sentence. However, such uniform reinforcement fails to distinguish critical, high-uncertainty decision points from certain ones, leading to inadequate updates for challenging tokens.

To refine this, we propose an uncertainty-aware policy optimization strategy that amplifies the optimization signal at high-uncertainty positions. First, for each candidate $Y^{(i)}$ in $\mathcal{S}$, we compute the sequence-level reward $R(Y^{(i)})$ against the constructed pseudo-label Y^* using BLEU:

$$R(Y) = \mathrm{BLEU}(Y^{(i)}, Y^*). \tag{13}$$

We then normalize these rewards within the group following the GRPO framework (Eq. 4) to derive the raw advantage $A_t^{(i)} = A^{(i)}$ for each token.

To identify critical optimization targets, we still utilize token-level entropy H_t as a proxy for model uncertainty. Drawing on the observation that effective RL is driven by a minority of high-entropy tokens (Wang et al., 2025b), we generate a binary importance mask $M^{(i)} \in \{0,1\}^L$ for each sequence of length L. The mask $M_t^{(i)}$ selects tokens falling into the top-q fraction of entropy within the sequence, defined as:

$$M_t^{(i)} = \begin{cases} 1, & \text{if } H_t > \mathcal{Q}_q(\{H_j\}_{j=1}^L) \\ 0, & \text{otherwise} \end{cases}, \tag{14}$$

where $\mathcal{Q}_q$ denotes the $(1-q)$-th quantile threshold of the entropy distribution for the current sentence. This dynamic thresholding ensures that high uncertainty is defined relatively to the complexity of the specific sentence context.

Finally, the advantage is reweighted to increase the contribution of these uncertain positions:

$$\hat{A}_t^{(i)} = A_t^{(i)} \cdot (1 + \alpha \cdot M_t^{(i)}), \tag{15}$$

where α is a hyperparameter controlling the amplification intensity. This selective reweighting intensifies feedback at ambiguous decoding steps. Under high uncertainty, tokens with positive advantages receive stronger reinforcement, while negative ones incur larger penalties. Concentrating optimization on these uncertain positions enables targeted refinement of challenging semantic decisions.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments primarily on bidirectional Chinese⇔English (ZH⇔EN) Task. Following Feng et al. (2025), we train on 13,130 high-quality in-distribution (ID) sentences from WMT17–20 and evaluate on WMT24 EN⇒ZH and WMT23 ZH⇒EN. For out-of-distribution (OOD) language generalization, we

Type	Method	Seen Domains												Unseen Domain			
		IT			Law			Medical			Subtitles			Koran			
		Bl.	Co.	Kw.	Bl.	Co.	Kw.	Bl.	Co.	Kw.	Bl.	Co.	Kw.	Bl.	Co.	Kw.	**Av.**
								Qwen2.5-0.5B-Instruct									
-	Base	28.38	76.23	54.63	26.18	76.36	54.59	17.14	74.36	53.67	18.20	72.99	53.37	7.61	60.26	46.95	48.06
Sup.	GRPO	28.79	76.09	**59.96**	26.31	76.42	**59.27**	25.32	74.09	55.32	**26.47**	74.13	55.24	7.93	59.25	44.39	49.93
Sup.	DAPO	29.25	77.22	58.30	18.91	74.29	55.44	**28.03**	**76.74**	**60.06**	18.17	72.02	**59.28**	8.46	61.30	47.23	49.65
Unsup.	TTRL	27.64	76.48	50.72	23.49	76.11	51.10	15.10	72.94	44.33	15.13	72.39	44.40	6.49	61.09	42.63	45.36
Unsup.	**Ours**	**30.23**	**77.65**	56.94	**27.44**	**77.37**	56.90	22.67	**75.82**	56.97	22.67	**75.82**	56.60	**8.53**	**61.51**	**47.51**	**50.31**
								Qwen3-4B-Instruct									
-	Base	35.72	81.45	64.44	34.09	83.69	70.77	38.09	82.10	68.20	25.20	77.07	70.23	14.62	71.43	64.96	58.80
Sup.	GRPO	**37.45**	82.65	65.24	**37.49**	84.45	**71.99**	**40.42**	83.44	70.27	**26.97**	**79.27**	71.29	15.20	72.04	65.54	**60.24**
Sup.	DAPO	37.03	82.11	64.98	36.69	83.69	70.11	38.90	83.10	69.86	26.10	78.96	71.17	14.99	71.89	65.46	59.67
Unsup.	TTRL	35.88	81.74	64.66	33.65	83.21	70.68	38.41	82.75	68.99	24.98	76.88	70.04	14.51	71.22	64.91	58.83
Unsup.	**Ours**	36.60	**82.80**	**65.25**	35.00	**84.53**	70.94	39.93	**83.91**	**70.52**	26.11	79.06	**71.62**	**15.43**	**72.62**	**65.84**	60.01
								Tower-Plus-2B									
-	Base	34.10	80.62	65.52	40.73	84.78	71.78	42.10	83.63	70.79	26.75	79.05	71.70	15.10	72.29	65.81	60.31
Sup.	GRPO	**38.44**	82.48	65.53	**47.97**	**86.12**	71.22	**45.40**	84.30	70.22	**28.36**	79.43	**71.88**	**15.74**	72.47	65.58	**61.60**
Sup.	DAPO	37.18	81.43	64.24	46.82	85.55	71.25	44.82	83.82	70.59	26.92	78.65	71.04	15.59	72.40	65.24	61.03
Unsup.	TTRL	35.79	81.41	65.04	40.87	84.72	**71.81**	42.08	83.51	70.80	26.40	78.82	71.38	14.66	71.92	65.79	60.33
Unsup.	**Ours**	37.24	**82.53**	**65.71**	43.04	85.92	71.62	42.92	**84.67**	**70.98**	27.39	79.25	71.46	15.16	**72.50**	**66.12**	61.13

Table 2: Domain Translation performance for DE⇒EN using BLEU (Bl.), XCOMET (Co.) and COMETKiwi (Kw.) metrics, with average metric scores (Av.). Supervised (Sup.) and Unsupervised (Unsup.) labels indicate the learning setting. The **best** results are bolded and the second best results are underlined.

test on WMT24 English⇒Japanese (EN⇒JA), WMT23 German⇒English (DE⇒EN), and Flores-200 German⇒Chinese (DE⇒ZH). We further assess robustness under domain shift using multi-domain German⇒English datasets (Aharoni and Goldberg, 2020), training on IT, Law, Medical, and Subtitles domains, with Koran as an unseen domain. Detailed statistics are in Appendix A.

Backbone Models. We employ three instruction-tuned LLMs with varying scales: Qwen2.5-0.5B-Instruct (Qwen et al., 2025), Qwen3-4B-Instruct (Yang et al., 2025a), and Tower-Plus-2B (Rei et al., 2025) which is specifically optimized for multilingual MT tasks.

Baselines. We compare with two supervised baselines GRPO and DAPO following Feng et al. (2025), and an unsupervised baseline TTRL.

Evaluation Metrics. We assess performance using *SacreBLEU* (Post, 2018) for lexical metric, and two semantic metrics: reference-based XCOMET-XL (Guerreiro et al., 2024) and reference-free COMETKiwi-23-XL (Rei et al., 2022).

Training Details Following (Feng et al., 2025), we implemented our framework based on verl[1]. More details are in Appendix C.

4.2 Translation Results

Overall Performance. Table 1 reports results on five translation directions. Our method achieves the best average score of **57.29**, outperforming supervised RL (avg. **56.66**) and unsupervised TTRL (avg. **54.77**) by **+0.63** and **+2.52 points** respectively, demonstrating the effectiveness of our entropy-driven RL for unsupervised MT.

Compared with the unsupervised baseline TTRL, our method achieves substantial improvements in all settings. In contrast, TTRL often shows limited gains or even degradation due to consensus bias and mode collapse. Relative to supervised RL methods (GRPO and DAPO), our framework surpasses them in average scores and semantic metrics (Co./Kw.). Although GRPO occasionally attains higher BLEU on specific directions, this stems from that supervised RL is reference-driven optimization, which favors exact lexical and n-gram matching for BLEU metric. Our method encourages exploration of pseudo-labels prioritizing semantic faithfulness and overall adequacy over strict word-level alignment. Furthermore, our method shows strong generalization on OOD language pairs, suggesting that entropy-driven RL facilitates learning universal translation patterns rather than overfitting to ID languages.

4.3 Domain Translation Results

Table 2 shows results across domains. Our method consistently achieves the highest average scores in the unsupervised setting, with notable gains in semantically diverse domains like IT and Subtitles. However, a performance gap persists compared to supervised GRPO in highly specialized domains (e.g., Medical), particularly in BLEU. We attribute this to the *knowledge ceiling* of self-evolving RL:

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

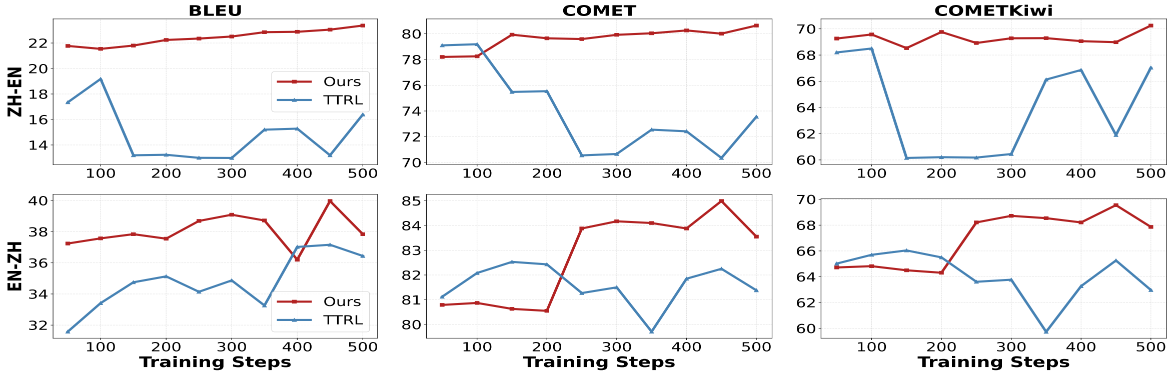


Figure 3: Performance over training steps on Qwen3-4B-Instruct. Our method achieves higher and more stable improving trend throughout training, while TTRL exhibits large fluctuations and periodic performance drop.

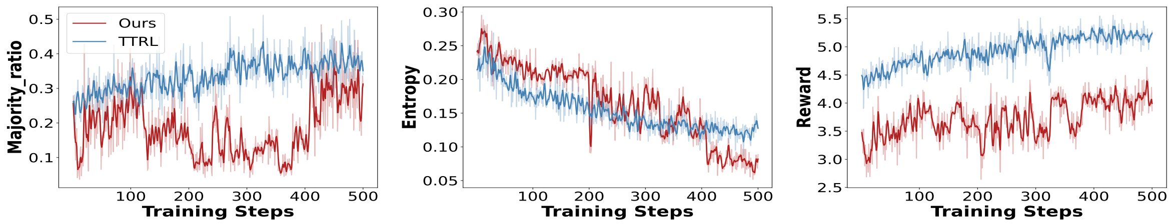


Figure 4: Training dynamics on Qwen3-4B-Instruct ZH⇔EN. Our method maintains higher entropy and lower early-stage consensus, preventing premature collapse, while TTRL quickly converges to consensus states; consequently, our reward improves more steadily and stabilizes over training.

As noted by Zuo et al. (2025) and Yue et al. (2025), unsupervised RL success depends on the model's prior knowledge, and while RL improves sampling efficiency toward the correct path, it cannot generate precise supervision for rare terminologies absent from the backbone's knowledge base, nor elicit fundamentally new reasoning patterns. Instead, supervised RL benefits from references that provide direct domain-specific guidance.

4.4 Performance over Training Steps

Figure 3 visualizes training stability over 500 steps on ZH⇔EN with Qwen3-4B-Instruct. Unlike the performance collapses of TTRL, our method maintains a stable ascent trend. Although both employ Test-Time Scaling (TTS) (Snell et al., 2025) via majority voting, TTRL triggers consensus bias, causing large fluctuations and performance drop. This indicates that simply scaling the number of samples is insufficient for unsupervised MT. Our method mitigates this through Entropy-based Dynamic Sampling and Confidence-Weighted Voting. The former increases information gain during K-sample generation, while the latter converts TTS into reliable training signals. Finally, by focusing optimization on critical decision points, we prevent the model from converging toward local optima.

4.5 Training Dynamics for Unsupervised RL

Figure 4 plots the majority ratio, entropy and reward during training. The majority ratio measures the proportion of samples in $\mathcal{S}$ that match the pseudo-label Y^* , reflecting sampling consensus. TTRL exhibits a faster increase in both majority ratio and reward. However, this trend leads to a phenomenon: homogeneous samples dominate and are further reinforced by majority voting, forming a consensus bias loop. Conversely, our method maintains higher entropy and a lower majority ratio during early training, preserving translation diversity and enabling sustained exploration. As uncertainty is progressively resolved, the majority ratio increases and rewards improve steadily. This behavior indicates that our framework leverages TTS for semantic exploration rather than premature exploitation, achieving more robust self-evolution.

4.6 Impact of Uncertainty Selection Ratio q

Figure 5 illustrates the impact of the uncertainty selection ratio q in Eq. 14 on training stability. We find that $q = 20\%$ achieves the best performance, confirming the finding of Wang et al. (2025b) that focusing on the top 20% of high-entropy tokens facilitates effective and stable RL training. Increasing

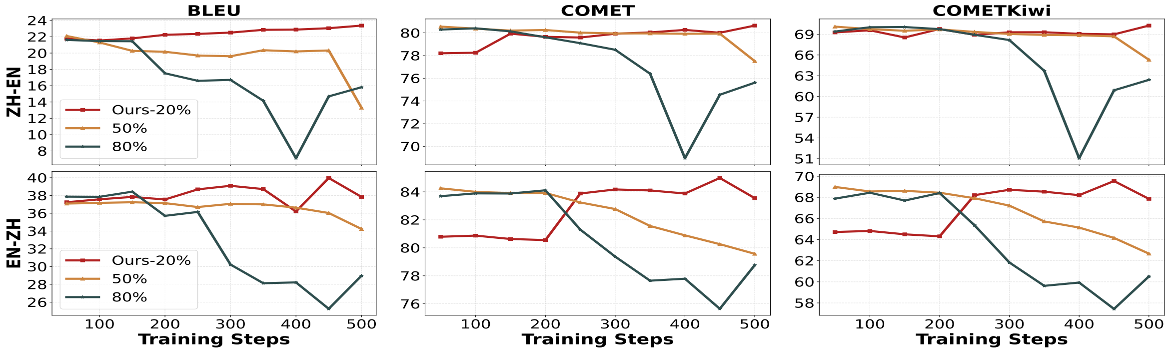


Figure 5: Performance over training steps under different entropy selection ratio $q \in \{20\%, 50\%, 80\%\}$ on Qwen3-4B-Instruct ZH⇔EN, where only the top-q high-entropy tokens per sentence receive amplified policy updates.

q to 50% or 80% leads to significant fluctuations and eventual performance collapse. These results suggest that including excessive low-entropy tokens dilutes the reward signal with noise, whereas targeting high-uncertainty pivotal tokens is essential for maintaining robust self-evolving learning.

4.7 Ablation Study

Method	ZH⇒EN	EN⇒ZH
Ours	16.07	27.15
w/o EDTS	14.13	26.15
w/o CW-SRLG	13.66	25.64
w/o UAPO	14.83	25.80

Table 3: Ablation study and report BLEU scores in ZH⇔EN with Qwen2.5-0.5B-Instruct.

We conduct ablation experiments to assess the contribution of each core component in our framework. *w/o* EDTS uses a fixed sampling temperature, *w/o* CW-SRLG generates pseudo-labels via vanilla majority voting, and *w/o* UAPO assigns equal advantage weights to all tokens. As shown in Table 3, removing any component consistently degrades performance. In particular, removing EDTS reduces exploration over diverse yet valid translations, removing CW-SRLG increases the risk of consensus bias and causes the largest performance drop, and removing UAPO weakens learning on high-uncertainty tokens. These results demonstrate that the three components work synergistically to improve translation quality.

4.8 Generalization to Low-Resource Languages

To evaluate the broader applicability of our framework in low-resource settings, we further extend our experiments on Qwen3-4B-Instruct to two representative language pairs from Flores-200 (Team et al., 2022): English⇒Swahili (an African Bantu language) and English⇒Lao (a Southeast Asian Kra-Dai language). Since the Flores corpus does not provide an official training split, we use the dev set (997 sentences) for RL fine-tuning and the test set (1012 sentences) for evaluation. In addition to standard MT metrics, we also reported chrF (Popović, 2015), which provides a more reliable character-level evaluation for morphologically rich or low-resource languages. Results are reported in Table 4.

Method	Bl.	ch.	Co.	Kw.
	English⇒Swahili			
Base	2.91	9.48	37.68	8.87
GRPO$_{sup.}$	4.65	12.73	40.55	11.45
TTRL$_{unsup.}$	2.88	9.41	37.21	8.76
Ours$_{unsup.}$	**4.74**	**12.87**	**41.02**	**11.78**
	English⇒Lao			
Base	6.68	14.15	42.86	17.51
GRPO$_{sup.}$	8.98	16.68	45.67	20.23
TTRL$_{unsup.}$	8.07	15.66	44.45	19.41
Ours$_{unsup.}$	**9.26**	**16.87**	**45.99**	**20.75**

Table 4: Low-resource language MT experiments with Qwen3-4B-Instruct using BLEU (Bl.), chrF (ch.), XCOMET (Co.) and COMETKiwi (Kw.) metrics. Supervised (*sup.*) and Unsupervised (*unsup.*) labels indicate the learning setting.

Our method consistently outperforms GRPO and TTRL across all evaluation metrics for both language pairs. This confirms that our unsupervised RL paradigm can effectively optimize translation policies even in linguistically diverse and data-scarce scenarios. Notably, on English⇒Swahili, TTRL slightly degrades performance compared to the Base model (2.91→2.88 BLEU). This behavior empirically supports our consensus bias hypothesis: when training data is extremely limited, TTRL's reliance on model-generated consensus can amplify spurious agreements, leading to self-reinforcing errors. In contrast, our entropy-driven mechanism prioritizes high-uncertainty tokens, encouraging exploration beyond biased consensus. This is particularly beneficial in low-resource scenarios, where limited initial translation capability and scarce supervision make uncertainty inherently higher.

5 Related Work

RL for LLM-based MT. Recent advances integrate reasoning into MT-RL. R1-T1 (He et al., 2025) leverages Chain-of-Thought reasoning with hybrid rewards to incentivize human-like inference during translation. MT-R1-Zero (Feng et al., 2025) adapts the R1-Zero paradigm, applying large-scale RL directly on pretrained LLMs using reference-based and reference-free rewards. DeepTrans (Wang et al., 2025a) focuses on deep reasoning for literary and free translation via RL with a process-aware reward model. In specialized domains, TAT-R1 (Li et al., 2025b) enhances terminology accuracy by combining RL with word alignment techniques. Other efforts focus on reward stability and autonomy: Tan and Monz (2025) utilize human preference models to enhance robustness, RIVAL (Li et al., 2025a) employs adversarial min-max framework for stable RL updates, and SSR-Zero (Yang et al., 2025b) introduces self-rewarding mechanisms to eliminate external dependencies.

Entropy-Based RL in LLMs. Entropy is a pivotal indicator of uncertainty for guiding RL. Wang et al. (2025b) and Liu et al. (2025) demonstrate that RL gains concentrate on high-entropy "forking tokens", utilizing entropy-fork trees to balance exploration and exploitation. To enhance training, EDGE-GRPO (Zhang et al., 2025) implements entropy-guided advantage diversification for gradient quality, while Hao et al. (2025) propose adaptive reweighting to prevent premature convergence.

Existing MT-RL relies on supervised rewards, while entropy-based RL is confined to reasoning tasks. We bridge this gap by enabling self-evolving MT via entropy without external supervision.

6 Conclusion

In this paper, we proposed an entropy-driven unsupervised RL framework for MT, which uses entropy as an uncertainty proxy, enabling reward signals to co-evolve with model competence. Through entropy-guided sampling, confidence-weighted labeling, and uncertainty-aware optimization, our approach mitigates consensus bias and promotes semantic refinement without fixed references. Empirical results across multiple high- and low-resource languages and domains demonstrate significant gains over both supervised and unsupervised baselines, highlighting entropy-driven self-evolution as a scalable direction for advancing MT.

Limitations

Our experiments have not evaluated the method's applicability to other research areas, such as multimodal translation. In future work, we plan to extend our unsupervised RL framework for MT to support a broader range of applications.

Ethics Statement

This research adheres to a strict ethical framework as it does not involve any ethical issues. The data constructed for this research is derived solely from open-source data, and the large language models employed in this study follows their declared licenses.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. U24A20334, 62366027, U21B2027, 62266027), the Yunnan Provincial Major Science and Technology Special Plan Projects (Grant Nos. 202303AP140008, 202203AA080004, 202302AD080003, 202401BC070021) and the Science and Technology Projects of Yunnan Universities Serving Key Industries (Grant No.FWCY-ZD2025006).

References

Roee 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.

Julius Cheng and Andreas Vlachos. 2024a. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128.

Julius Cheng and Andreas Vlachos. 2024b. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Jian Wu, and Zuozhu Liu. 2025. MT-r1-zero: Advancing LLM-based machine translation via r1-zero-like reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18685–18702, Suzhou, China. Association for Computational Linguistics.

Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. 2024. Rlef: Grounding code llms in execution feedback with reinforcement learning. *arXiv preprint arXiv:2410.02089*.

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

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*.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. 2025. Rethinking entropy interventions in rlvr: An entropy change perspective. *arXiv preprint arXiv:2510.10150*.

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.

Shuting Jiang, Ran Song, Yuxin Huang, Yan Xiang, Yantuan Xian, Shengxiang Gao, and Zhengtao Yu. 2026. Consensus-aligned neuron efficient fine-tuning large language models for multi-domain machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31310–31318.

Yoko Kayano and Saku Sugawara. 2025. Specification-aware machine translation and evaluation for purpose alignment. In *Proceedings of the Tenth Conference on Machine Translation*, pages 113–141.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, H Ivison, F Brahman, LJV Miranda, A Liu, N Dziri, S Lyu, and 1 others. 2025. Tulu 3: Pushing frontiers in open language model post-training, 2024. *URL https://arxiv.org/abs/2411.15124*, 297.

Long Li, Xuzheng He, Haozhe Wang, Linlin Wang, and Liang He. 2024. How do humans write code? large models do it the same way too. *arXiv preprint arXiv:2402.15729*.

Tianjiao Li, Mengran Yu, Chenyu Shi, Yanjun Zhao, Xiaojing Liu, Qi Zhang, Xuanjing Huang, Qiang Zhang, and Jiayin Wang. 2025a. RIVAL: Reinforcement learning with iterative and adversarial optimization for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3064–3079, Suzhou, China. Association for Computational Linguistics.

Zheng Li, Mao Zheng, Mingyang Song, and Wenjie Yang. 2025b. Tat-r1: Terminology-aware translation with reinforcement learning and word alignment. *Preprint*, arXiv:2505.21172.

Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. 2025. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. *arXiv preprint arXiv:2508.11356*.

Yilun Liu, Shimin Tao, Chang Su, Min Zhang, Yanqing Zhao, and Hao Yang. 2022. Part represents whole: Improving the evaluation of machine translation system using entropy enhanced metrics. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 296–307, Online only. Association for Computational Linguistics.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*.

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

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

Qwen, :, 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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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

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.

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*.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Ran Song, Shengxiang Gao, Xiaofei Gao, Cunli Mao, and Zhengtao Yu. 2025. Mke-pllm: A benchmark for multilingual knowledge editing on pretrained large language model. *Neurocomputing*, 651:130979.

Shaomu Tan and Christof Monz. 2025. Remedy: Learning machine translation evaluation from human preferences with reward modeling. *arXiv preprint arXiv:2504.13630*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, 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. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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, and Jie Zhou. 2025a. Deep reasoning translation via reinforcement learning. *arXiv preprint arXiv:2504.10187*.

Junqiao Wang, Zeng Zhang, Yangfan He, Zihao Zhang, Xinyuan Song, Yuyang Song, Tianyu Shi, Yuchen Li, Hengyuan Xu, Kunyu Wu, and 1 others. 2024a. Enhancing code llms with reinforcement learning in code generation: A survey. *arXiv preprint arXiv:2412.20367*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. 2024b. Reinforcement learning enhanced llms: A survey. *arXiv preprint arXiv:2412.10400*.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. 2025. Self-rewarding correction for mathematical reasoning. *arXiv preprint arXiv:2502.19613*.

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

Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, and Sitong Wang. 2025b. Ssr-zero: Simple self-rewarding reinforcement learning for machine translation. *Preprint*, arXiv:2505.16637.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, and 1 others. 2025. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.

A Data Statistics

This section provides further details on the datasets used in our experiments for both ZH⇔EN MT task and DE⇒EN domain MT task. For ZH⇔EN MT task, as shown in Table 5, we use the same dataset in (Feng et al., 2025), containing 13,130 ZH⇔EN parallel sentence pairs. For supervised RL baselines, we use the bilingual data, and for unsupervised RL baseline and our method, we only use the source text for training. For DE⇒EN domain MT task, we use multi-domain datasets from Tian et al. (2014). We randomly select 1k samples of each domain for training and use the standard test set for evaluation. Table 6 is detailed statistics.

B Translation Prompts

We show our translation prompts below:

ZH=>EN Translation Prompts

System: You are a helpful translation assistant. Translate the following sentence from {src_language} into {tgt_language} and only reply the translated sentence without line breaks or any special characters.\nUser: {src_text}\nAssistant:

EN=>ZH Translation Prompts

System:你是一位翻译助手。请将以下句子从英语翻译成中文，并只回复翻译结果，不要生成换行或任何特殊符号。\nUser: {src_text}\nAssistant:

Here, {src_language} and {tgt_language} is the source and target language in MT, and {src_text} is the source text to be translated. To ensure the generated outputs are compatible with automated evaluation metrics, we explicitly instruct the model to output only the translation without any line breaks or special characters. Furthermore, we employ target-language specific system instructions to better prime the model into the desired linguistic context and minimizes potential ambiguities.

C Implementation Details

During training, we configure a global training batch size of 32. We sample $K = 64$ rollouts for each training prompt within the GRPO algorithm. We employ a constant learning rate of 5e-7. For the **EDTS** module, we set the baseline temperature $T_{\text{base}} = 1.0$ and the scaling factor $\lambda = 0.05$ in Eq. 5. For the **UAPO** module, the entropy selection ratio q in Eq. 14 is set to 0.2, and the amplify intensity α in Eq. 15 is set to 1.0. We tested $\alpha \in [1.0, 2.0]$ and found $\alpha = 1.0$ provides the most stable and consistent improvements across language pairs. The maximum generation length for responses is 256 tokens. Regarding optimization, we set the GRPO clipping threshold $\varepsilon = 0.2$ and remove the KL penalty ($\beta = 0$), as it tends to restrict the policy from fully exploring diverse responses. All models are trained for 2 epochs on 8 NVIDIA A40 GPUs.

D Bootstrap significance tests

To demonstrate that the observed improvements are not due to chance, we conducted two levels of bootstrap significance tests: **Statistical Significance** (Test Phase) and **Training Stability** (Training Phase). These tests provide a robust evaluation of both the statistical reliability of our gains and the stability of our training process.

Statistical Significance (Test Phase). We performed Bootstrap Resampling with $N = 1000$ iterations on the BLEU, COMET, and COMETKiwi scores for both the EN⇒ZH and ZH⇒EN tasks. The p-values obtained from these tests are summarized in Table 7. Our method consistently outperforms both the GRPO and TTRL baselines, with p-values well below the 0.05 threshold, and many even reaching $p < 0.001$. This strong statistical significance provides compelling evidence that the improvements are not incidental but are robust and reliable across different evaluation metrics (surface-level accuracy and semantic adequacy).

Training Stability (Training Phase). To validate the training stability of our approach, we retrained Qwen3-4B-Instruct using three different random seeds (Seed 1, 42, and 3407) and evaluated its performance across the same set of metrics. As shown in Table 8, the performance of our method demonstrates extremely low variance across the different random seeds, with the average BLEU score across the seeds remaining consistent within a narrow range. This stability indicates that our entropy-driven unsupervised RL framework is not only effective but also robust to initialization variations, ensuring that the observed improvements are reliable throughout the training process.

E Comparison of Computational Costs

To ensure a fair comparison, we maintain identical training configurations (e.g., rollout counts, batch sizes, and model parameters) across our method and all baselines (GRPO, DAPO, and TTRL). Computational costs for the Qwen3-4B-Instruct model

	Train		Test				
	EN⇒ZH	ZH⇒EN	EN⇒ZH	ZH⇒EN	EN⇒JA	DE⇒EN	DE⇒ZH
Source	WMT 17-20		WMT 24	WMT 23	WMT 24	WMT 23	Flores
statistics	6565	6565	997	1976	997	549	1012

Table 5: Data statistics for the training and test sets used in the main ZH⇔EN translation experiments.

	IT	Law	Medical	Subtitles	Koran
Train	1000	1000	1000	1000	-
Test	2000	2000	2000	2000	2000

Table 6: Data statistics for the training and test sets used in DE⇒EN domain translation experiments.

are shown in Table 9; notably, our approach introduces minimal computational latency compared to the baselines. Specifically, our entropy-based modules, dynamic temperature scaling, and advantage reweighting are computed directly from existing token distributions or through simple logit/loss scaling, requiring no additional forward or backward passes. Furthermore, as our approach aligns with the Test-Time Scaling (TTS) paradigm, where performance improves with more rollouts and optimization steps, we achieve superior performance while maintaining a training wall-clock time nearly identical to that of GRPO and TTRL.

F High and Low Entropy Token Distributions

To validate the entropy reflects translation difficulty, we visualize token entropy distribution for the DE⇒EN task across three domains in Figure 6. High-entropy tokens (upper row) correspond to polysemous, context-dependent words that require semantic disambiguation: **IT Domain** (Figure 6a): Terms like "set" and "filter" exhibit high entropy since their translation depends strictly on their grammatical role (noun vs. verb) and technical context (e.g., "a set of data" vs. "to set a variable"). **Medical Domain** (Figure 6b): The token "drug" can refer to general medication, specific narcotics, or clinical compounds depending on the therapeutic setting. Similarly, "transport" may involve biological protein transport or the physical logistics of medical supplies. **Law Domain** (Figure 6c): Keywords such as "subject" and "relevant" are highly ambiguous; "subject" can denote a legal entity, a topic of a contract, or the act of being "subject to" a regulation.

In contrast, low-entropy tokens (dates, numbers, formatting items), and frequent syntactic markers (56th, 2012/03, Bände, Herz) exhibit stable cross-lingual mappings, leading to almost deterministic model decisions. Overall, uncertainty concentrates on domain-sensitive terminology and polysemous content words, while structural tokens remain low-entropy. Our method identifies reliable pseudo-labels among these uncertain variants, enabling the model to resolve complex semantic ambiguities.

G Case Study

Figure 7 qualitatively compares the translation evolution of vanilla TTRL and our framework over 500 training steps. TTRL exhibits noticeable instability and inconsistency in lexical choices and sentence formulation. For instance, it fluctuates between different renderings of “adapt” and consistently adopts the literal but less professional translation “决策体” for “decision-making body”. Notably, this suboptimal choice persists unchanged from Step 50 to Step 500, indicating premature convergence driven by unweighted majority voting. As training progresses, TTRL’s outputs also become increasingly verbose and repetitive, with rigid sentence structures and awkward handling of complex clauses, such as the parenthetical description of suburban councils at Step 500. These qualitative patterns align with the performance fluctuations observed in our quantitative analysis.

Conversely, our framework demonstrates a stable and progressive refinement process. At the lexical level, it resolves domain-specific entities more accurately, for example expanding “NSW” into the formal Chinese name “新南威尔士州”, and replaces literal expressions with more appro-

Baseline	ZH⇒EN			EN⇒ZH		
	BLEU	COMET	COMETKiwi	BLEU	COMET	COMETKiwi
vs. GRPO	0.0009	0.0023	< 0.0001	0.0019	0.0430	< 0.0001
vs. TTRL	0.0009	< 0.0001	< 0.0001	0.0089	< 0.0001	< 0.0001

Table 7: p-values from Bootstrap Significance Tests comparing our method with GRPO and TTRL on ZH⇔EN using Qwen3-4B-Instruct.

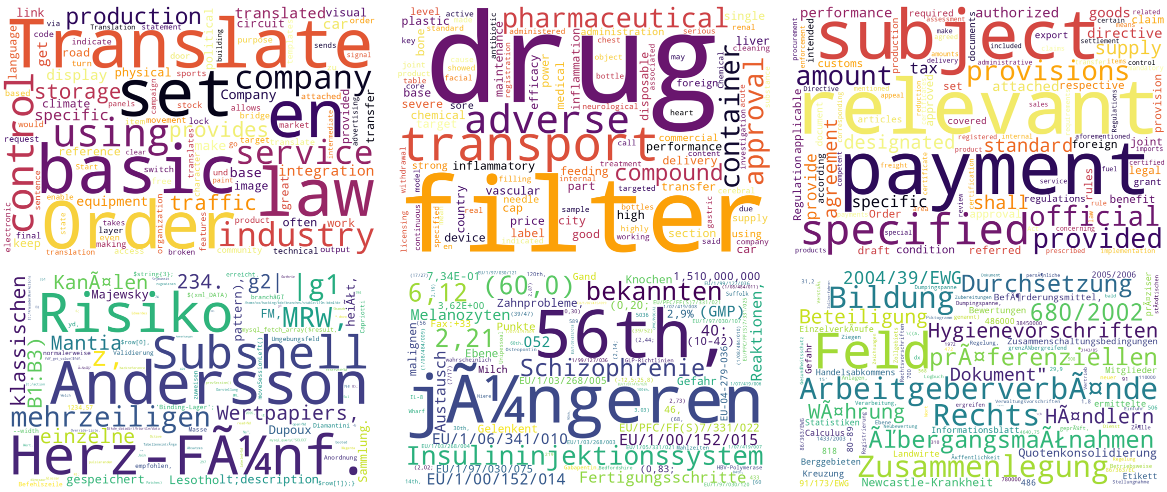
Seed	ZH⇒EN			EN⇒ZH		
	BLEU	COMET	COMETKiwi	BLEU	COMET	COMETKiwi
1	23.37	80.64	70.25	39.97	84.99	69.55
42	23.54	81.03	70.62	40.12	85.06	69.60
3407	23.16	79.94	69.86	39.64	84.51	69.02
Avg ± Std	23.36±0.19	80.54±0.55	70.24±0.38	39.91±0.25	84.85±0.30	69.39±0.32

Table 8: Training stability across different random seeds on ZH⇔EN with Qwen3-4B-Instruct.

Method	Training Wall-Clock Time (Hours)
GRPO	40.98
DAPO	41.26
TTRL	41.44
Ours	42.05

Table 9: Training Wall-Clock Time for each method.

priate professional terms such as “决策机构” and “未来面貌”. Structurally, our method gradually improves its handling of long and nested clauses, producing more coherent sentence organization and better integration of descriptive modifiers. By Step 500, the model generates a polished translation that faithfully captures both the semantic content and institutional tone of the source text. This qualitative progression confirms that by mitigating consensus bias and emphasizing high-uncertainty tokens, our framework successfully internalizes Test-Time Scaling gains into the model parameters, yielding translations that are both semantically adequate and stylistically professional.



(a) IT Domain (b) Medical Domain (c) Law Domain

Figure 6: Word clouds of the top 100 tokens with the highest (upper) and lowest (lower) average entropy for different domains, respectively, selected from the set of frequently occurring tokens. High-entropy tokens align with domain-specific keywords, while low-entropy tokens correspond to deterministic entities or syntactic structures.

EN=>ZH Source Text

User: Our planning system also needs to adapt so that planners in suburban councils, who seem to be less supportive of change, are not the only determinants of the future shape of their suburbs. A way forward could be to have an independent decision-making body that championed quality design as long as it fitted below the tree tops and incorporated family living, large balconies, common gardens and presented a friendly appearance. This could be a role for the current NSW government architect working within a carefully crafted complying code.

TTRL Step 50

Assistant: 我们的**规划系统**也需**适配**以便郊区议会的规划师们似乎对变革支持较少不是决定其郊区**未来形状**的唯一因素。**一个可行的路**是设立一个独立的**决策体长期倡导**高质量的设计只要设计不超出树冠线并融入家庭生活大阳台共同花园和友好的外观。这由**目前NSW**政府建筑师在精心设计的**合规代码**内承担。

TTRL Step 250

Assistant: 我们的**规划体系**也需**作出调整**以便郊区议会的人员尽管似乎对变革支持较少却不再成为其郊区**未来发展形状**的唯一决定因素。**一个可行的路**是设立一个独立的**决策体**应**致力于**推动高质量的设计只要设计不超出树冠线融入家庭生活大阳台共享花园以及友好的外观即可。这一角色可由**目前NSW**政府建筑师在精心设计的**合规代码**内承担。

TTRL Step 500

Assistant: 我们的**规划系统**也需要**适应**，以便使郊区议会的规划师——他们似乎对变化的支持较少——不是决定其郊区**未来形状**的唯一因素。**一个可行的路**是设立一个独立的**决策体**，**致力于**高质量的设计，只要设计不超出树冠线并融入家庭生活大阳台共享花园以及友好的外观即可。这个角色可以由**目前NSW**政府建筑师在精心设计的**合规代码**内承担。

Ours Step 50

Assistant: 我们的**规划系统**也需**进行调整**，以确保郊区议会的规划人员——他们似乎对变革支持较低——并非决定其郊区未来**形态**的唯一因素。**一个可行的方向**是设立一个独立的**决策机构**，该机构长期倡导高质量的设计，只要设计不超出树冠线，同时融入家庭生活、宽敞阳台、共享花园，并呈现出友好的外观。这一角色可以由**目前新南威尔士州**政府建筑师在精心设计的**合规性规范框架内**承担。

Ours Step 250

Assistant: 我们的**规划体系**也需**做出调整**，以便让那些对变革不太支持的郊区议会规划人员不再成为其郊区未来**面貌**的唯一决定因素。**一个可行的方案**是设立一个独立的**决策机构**，该机构长期倡导高质量的设计，只要设计不超出树冠线，并融入家庭生活、宽敞的阳台、公共花园，以及友好的外观。这一角色可以由**新南威尔士州现任**政府建筑师在精心设计的**合规规范框架内**工作。

Ours Step 500

Assistant: 我们的**规划体系**也需**进行调整**，不能让那些似乎对变革不太支持的郊区议会规划人员成为决定郊区未来**面貌**的唯一因素。**一个可行的方案**是设立一个独立的**决策机构**，只要设计方案高度保持在树冠以下、融入适合家庭生活的设计、大阳台、公共花园以及友好的外观，该机构就应支持这些高质量的设计。这一角色可以由**现任新南威尔士州**政府建筑师在精心设计的**合规规范框架下**承担。

Figure 7: Comparison of translation evolution between TTRL and our framework over 500 steps on the EN⇒ZH task with Qwen3-4B-Instruct.