Incentivizing Parametric Knowledge via Reinforcement Learning with Verifiable Rewards for Cross-Cultural Entity Translation

Jiang Zhou[1], Xiaohu Zhao[2], Xinwei Wu[1], Tianyu Dong[1], Hao Wang[2], Yangyang Liu[2], Heng Liu[2], Linlong Xu[2], Longyue Wang[2], Weihua Luo[2], Deyi Xiong[1]†

[1]TJUNLP Lab, Tianjin University, China
[2]Alibaba Group, China
{jiangzhou, dyxiong}@tju.edu.cn

Abstract

Cross-cultural entity translation remains challenging for large language models (LLMs) as literal or phonetic renderings are usually yielded instead of culturally appropriate translations in context. However, relevant knowledge may already be encoded in model parameters during large-scale pre-training. To incentivize the effective use of parametric knowledge, we propose **EA-RLVR** (Entity-Anchored Reinforcement Learning with Verifiable Rewards), a training framework that optimizes cross-cultural entity translation without relying on external knowledge bases. EA-RLVR anchors supervision on a verifiable, entity-level reward signal and incorporates lightweight structural gates to stabilize optimization. This design steers the model toward learning a robust reasoning process rather than merely imitating reference translations. We evaluate EA-RLVR on XC-Translate and observe consistent improvements in both entity translation accuracy and out-of-domain generalization. Specifically, training on merely 7k samples boosts Qwen3-14B's entity translation accuracy from 23.66% to 31.87% on a 50k test set comprising **entirely unseen entities**. The learned entity translation ability also transfers to general translation, yielding +1.35 XCOMET on WMT24pp, which scales to +1.59 with extended optimization. Extensive analyses of $pass@k$ dynamics and reward formulations attribute these gains to superior sampling efficiency and a stable optimization landscape.

1 Introduction

At its core, machine translation aspires to make culturally situated texts accessible across languages. Despite substantial progress with multilingual large language models (Pan et al., 2025a), current systems often fall short of this goal in settings where translation hinges on culturally grounded entities

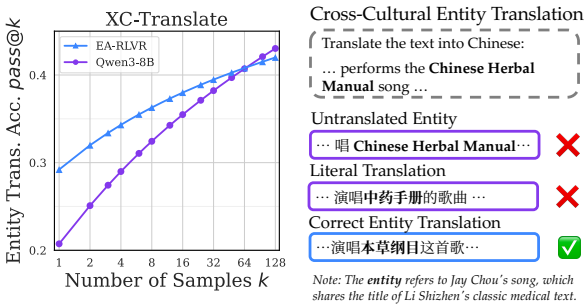

Figure 1: **(Left)** Entity translation accuracy (%) $pass@k$ curves demonstrate the base model possesses latent knowledge (high accuracy at large k) that EA-RLVR effectively activates at $k = 1$. **(Right)** An illustration of the challenge in cross-cultural entity translation.

such as books, films, places, songs and idioms (Yao et al., 2024). In these cases, producing an accurate, culture-aligned translation requires identifying, in context, which real-world entity is being referred to and how it is conventionally named in the target culture (Moghe et al., 2025). Recent evaluations have shown that even frontier proprietary LLMs frequently default to literal or phonetic renderings that are grammatically well formed but semantically inappropriate in context, thereby altering or obscuring the intended meaning of the source text (Conia et al., 2024).

A widely adopted workaround for this limitation is to equip translation systems with external knowledge, e.g., through online retrieval, knowledge graphs, or curated databases (Conia et al., 2024; Khandelwal et al.). These approaches can improve accuracy when relevant information is successfully retrieved. However, they also introduce practical and structural constraints. The performance of such systems depends critically on how well the task aligns with underlying database (Agrawal et al., 2023), and in practice often requires task-specific retrievers that must be trained or tuned (Wang et al., 2025b). Moreover, it fundamentally shifts the bottleneck from contextual entity reasoning to the

† Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5616–5638
July 2-7, 2026

structure and coverage of the external knowledge source, making translation quality contingent on what can be retrieved.

On the other hand, as trained on corpora spanning trillions of tokens across diverse domains and languages, LLMs implicitly encode a wide range of entity correspondences, cultural references, and real-world usage conventions (Yang et al., 2025; Qwen et al., 2025). In principle, such knowledge should support cross-cultural translation. As illustrated in Figure 1 (Left), the correct cultural entities are often present in the base model's probability distribution, evidenced by high accuracy when multiple sampling attempts ($pass@128$). However, such knowledge remains effectively inaccessible during standard single-pass generation ($pass@1$). Consequently, models frequently default to verbatim copying or literal renderings that obscure the intended meaning, such as retaining the source term or translating a song title literally as a medical manual (Figure 1, Right). These observations suggest that the core difficulty lies less in the availability of knowledge itself, but more in the absence of mechanisms that incentivize the model to surface that knowledge in a context-sensitive manner.

To incentivize LLMs to leverage their parametric knowledge effectively, we propose **EA-RLVR** (**E**ntity-**A**nchored RL with **V**erifiable **R**ewards), a framework for cross-cultural entity translation driven by fully rule-based, automatically verifiable reward. We cast cross-cultural entity translation as a sequence decision problem: given a source sentence, the model produces its own candidate translations, and a deterministic verifier evaluates whether the output expresses the correct target-culture entities. Rather than imitating reference translations, the model learns from verifiable rewards assigned to its own trajectories, reinforcing the reasoning that produces the correct entities. Concretely, EA-RLVR uses an **entity-matching reward** based on normalized substring matching between the predicted entity and the gold entity set. To stabilize optimization and reduce degenerate behaviors, we further introduce **structural gates** that modulate the reward according to lightweight output constraints (e.g., a prescribed reasoning format and translation length). This design avoids neural reward models that require additional computation and can be vulnerable to reward hacking in long-horizon RL, and it also addresses our empirical finding that neural metrics fails to provide supervision for culturally grounded entity choices. With these verifiable rewards and an efficient critic-free policy optimization recipe, EA-RLVR established a stable RLVR training framework for cross-cultural entity translation.

We conduct extensive experiments to evaluate EA-RLVR, yielding three key insights into its efficacy and underlying mechanisms: **(1) EA-RLVR incentivizes parametric knowledge.** Training on only 7k examples generalizes to a 50k test set whose entities are entirely unseen during training, improving entity translation accuracy by +8.21%–9.06% across different model scales. **(2) The learned strategy transfers beyond entity evaluation.** On WMT24pp, our models achieve improvements of XCOMET by +1.25–1.35 points, even though XCOMET is never used as supervision. When scaling training to the full dataset and extending optimization to 1,000 steps, the gains increase to +1.59–1.68 points. **(3) In-depth analyses clarify the dynamics of learning.** $pass@k$ evaluation, neural reward comparison, cross-lingual generalization, and examinations of reward-hacking behavior all point to the same pattern: our method improves sampling efficiency and induces stable, cross-cultural translation strategies, rather than encouraging memorization.

Our contributions are as follows: **(1)** We propose a novel framework for cross-cultural machine translation based on RLVR, showing that entity translation can be improved without access to external databases by directly incentivizing context-appropriate entity choices. **(2)** We empirically demonstrate that this approach improves entity translation accuracy and general translation quality across languages and model scales, including settings involving entirely unseen entities. **(3)** We provide in-depth analyses that explain the mechanism behind these improvements.

2 Related Work

Cross-Cultural and Entity-Centric Machine Translation Prior work addresses the challenge of culturally grounded entities largely through two avenues: external knowledge integration and targeted data augmentation. Retrieval-based methods explicitly ground translation in external sources, utilizing multilingual knowledge graphs (e.g., KG-MT; Conia et al., 2024) or document stores (e.g., RAGtrans; Wang et al., 2025b) to resolve entity ambiguities. While these approaches mitigate hallucinations, they introduce a dependency on the

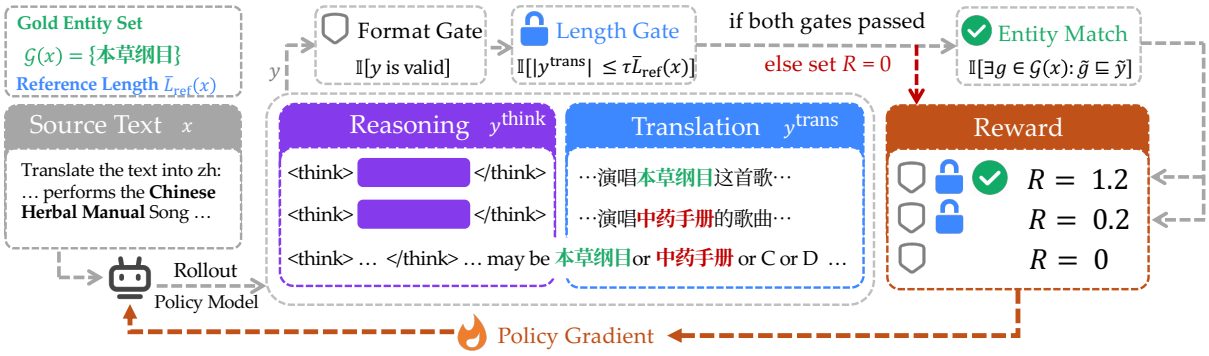

Figure 2: EA-RLVR framework: the policy model first rolls out a trajectory containing both reasoning and translation. This full trajectory must then pass two structural gates (format and length) to be eligible for a base reward (0.2 reward, $R = 0.2$). Finally, if the translation contains the correct entity, it receives an additional matching bonus (+1 reward, $R = 1.2$), and this final scalar reward drives the policy gradient update.

availability and quality of auxiliary databases. On the training side, recent works enhance entity robustness by synthesizing code-switched or entity-replaced data for denoising pre-training (Hu et al., 2022; Liang et al., 2024), or by jointly optimizing translation with entity alignment tasks (Riktters and Miwa, 2024). Language-aware parameter transfer methods further facilitate knowledge sharing across languages (Dong et al., 2025). Unlike these approaches, EA-RLVR does not require external retrieval at test time nor complex data synthesis pipelines. Instead, we cast entity translation as a reasoning problem, employing RLVR to activate and stabilize the parametric knowledge already present in the pre-trained model.

RLVR and Reasoning in Translation Recent post-training paradigms of LLMs leverage Reinforcement Learning with Verifiable Rewards (RLVR) to induce reasoning capabilities (Lambert et al., 2025; DeepSeek-AI et al., 2026), a process theoretically understood as improving sampling efficiency to activate latent knowledge already present in the base model (Yue et al., 2025; Huang et al., 2026; Dai et al., 2025; Yang et al., 2026; Jin et al., 2025). In machine translation, recent initiatives have actively explored integrating reasoning capabilities, for instance by employing multi-agent frameworks to synthesize long chain-of-thought trajectories for distillation (Wang et al., 2025a) or harnessing feedback from LLM judges and neural quality metrics to guide optimization (Feng et al., 2025a; Wang et al., 2025c; Feng et al., 2025b). Complementing these advances, EA-RLVR introduces a distinct paradigm centered on strict, rule-based verifiable rewards. We treat cultural entity translation as a precise reasoning task, employing deterministic rewards to directly surface parametric knowledge, thereby offering an alternative to distillation or neural-based objectives.

3 Method

We propose **EA-RLVR** (Entity-Anchored Reinforcement Learning with Verifiable Rewards), a framework designed to incentivize LLMs to accurately ground cultural entities during translation without external knowledge. As illustrated in Figure 2, our approach treats cross-cultural translation as a sequential decision process optimized via reinforcement learning.

As shown in Figure 2, the framework consists of three core components: (1) A **reasoning-aware policy** that generates a thinking trajectory before the final translation, allowing the model to elicit latent knowledge; (2) A **verifiable reward mechanism** that anchors supervision on deterministic entity matching, safeguarded by structural gates to prevent reward hacking; and (3) A **critic-free optimization algorithm** that stabilizes training using sequence-level importance ratio. In the following sections, we detail the task formulation (§3.1), the reward design (§3.2), and the policy optimization objective (§3.3).

3.1 Task Formulation

Given a source sentence x, we aim to generate a target-language translation y^{trans} that correctly renders the culturally grounded entity mention(s) in context. We treat an autoregressive LLM as a

stochastic policy π_θ over output tokens, i.e.,

$$\pi_\theta(y \mid x) = \prod_{t=1}^{|y|} \pi_\theta(y_t \mid x, y_{<t}).$$

Generation induces an episodic decision process in which the state at step t is $(x, y_{<t})$, the action is the next token y_t, and the episode terminates when an <eos> token is produced.

Reasoning and Translation Segments. Following recent reasoning-based post-training (DeepSeek-AI et al., 2026), the model is encouraged to produce a reasoning trace enclosed by <think> and </think> before emitting the final translation. When the response format is valid, we decompose the output y as

$$y = \langle \texttt{<think>} \rangle\ y^{\text{think}}\ \langle \texttt{</think>} \rangle\ y^{\text{trans}},$$

where y^{think} contains deliberation and y^{trans} is the final translation. Since the reasoning portion may contain exploratory candidate entities rather than the model's final decision, all content-based evaluation will be applied exclusively to y^{trans}.

3.2 Stable and Verifiable Reward Design for Cross-Cultural Entity Translation

Our main design goal is to construct a reward that is (i) *verifiable* from dataset annotations without a learned reward model, (ii) directly *aligned* with cross-cultural entity correctness, and (iii) *stable* under policy optimization and robust to reward hacking.

Normalized Entity Matching. Each example is annotated with a comprehensive set of acceptable target entities $\mathcal{G}(x)$, derived from the Wikidata alias field. This set captures legitimate variations, minimizing false negatives where the model predicts a valid entity surface form that differs from the primary reference. For brevity we denote $\tilde{y} = \text{norm}(y^{\text{trans}})$ and $\tilde{g} = \text{norm}(g)$, where $\text{norm}(\cdot)$ lowercases text and removes diacritics. We define a deterministic match function using normalized substring matching:

$$m(y, \mathcal{G}(x)) = \mathbb{I}\Big[\exists g \in \mathcal{G}(x) : \tilde{g} \sqsubseteq \tilde{y}\Big], \tag{1}$$

where $\mathbb{I}[\cdot]$ is the indicator function and $a \sqsubseteq b$ denotes that a is a substring of b. This captures whether the model produces an appropriate target-language realization of the entity.

Structural Gates: Format and Length. To ensure that rewards reflect meaningful translations rather than degenerate behaviors, we introduce two hard gates on the output. First, the response must follow the required <think> / </think> structure. Second, the translation segment must remain within a reasonable length relative to the references. Formally, let $g_{\text{fmt}}(y) \in \{0, 1\}$ indicate whether the format is valid, and define

$$g_{\text{len}}(x, y) = \mathbb{I}\Big[\ |y^{\text{trans}}| \le \tau \cdot \bar{L}_{\text{ref}}(x)\Big], \tag{2}$$

where $\bar{L}_{\text{ref}}(x)$ is the average reference length and $\tau > 0$ controls tolerance. Note that g_{len} is a *reward-side* hard gate: it zeroes out the reward for degenerate outputs but does not modify the policy gradient itself (cf. the *gradient-side* length normalization in Eq. 6). Only responses that satisfy *both* gates are eligible to receive any reward. We empirically demonstrate the necessity of these constraints in Appendix A, showing that removing them leads to catastrophic length explosion and reward hacking via keyword enumeration.

Final Reward. Combining these components, the terminal reward is

$$R(x, y) = g_{\text{fmt}}(y) g_{\text{len}}(x, y)\Big(\alpha + m(y, \mathcal{G}(x))\Big). \tag{3}$$

Intuitively, a response receives no reward if it fails either structural gate. If both gates are satisfied, it obtains a base reward α for producing a well-formed answer, and an additional bonus when the target entity is correctly realized. We set $\alpha = 0.2$ and $\tau = 2$ in all experiments.

3.3 Policy Optimization

Following recent advances in RL post-training for large language models, we optimize π_θ using a clipped policy-gradient objective from the PPO family (Schulman et al., 2017). To reduce training cost and improve stability, we adopt a critic-free variant and incorporate group-normalized advantages, sequence-level importance ratios, and asymmetric clipping, building on GRPO (Shao et al., 2024), GSPO (Zheng et al., 2025), and DAPO (Liu et al., 2025).

Objective. Given an input x, we sample G candidate responses $\{y_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$. The policy parameters are updated by maximizing

the clipped surrogate:

$$\mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\min\Big(s_i(\theta)\hat{A}_i,\right.$$
$$\left.\text{clip}\big(s_i(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\big)\hat{A}_i\Big)\right], \tag{4}$$

where $\mathcal{D}$ denotes the training dataset containing source sentences x, the ε_{low} and $\varepsilon_{\text{high}}$ are clipping thresholds that bound the policy update to prevent training instability, and the $s_i(\theta)$ is the sequence-level importance ratio defined in Eq. (6).

Group-normalized Advantages. Rather than relying on a learned critic, we compute an advantage for each sampled response relative to other responses in its *group* $\{y_i\}_{i=1}^{G}$:

$$\hat{A}_i = \frac{R(x,y_i) - \text{mean}(\{R(x,y_j)\}_{j=1}^{G})}{\text{std}(\{R(x,y_j)\}_{j=1}^{G})}, \tag{5}$$

where $\text{mean}(\cdot)$ and $\text{std}(\cdot)$ denote the sample mean and standard deviation within the group. This normalization stabilizes training and makes the reward scale largely irrelevant.

Sequence-level importance ratios. We compute the importance ratio at the sequence level with length normalization:

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)}\right)^{\frac{1}{|y_i|}}$$
$$= \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}\right). \tag{6}$$

This formulation discourages overly aggressive updates on long sequences while still allowing meaningful policy shifts when rewards are consistently better.

4 Experiments

Our experiments aim to verify two hypotheses: (1) that EA-RLVR training can effectively elicit latent cultural knowledge solely from the model's pre-trained parameters, and (2) that this entity-centric optimization does not compromise general translation quality. After outlining our setup in §4.1, we present empirical evidence supporting the activation of parametric knowledge in §4.2 and demonstrate positive transfer effects to general translation in §4.3.

4.1 Experimental Setup

Datasets and Benchmarks. To evaluate the activation of cultural knowledge, we utilized **XC-Translate** (Conia et al., 2024), a benchmark specializing in cross-cultural entity translation. We trained our models using 7,278 examples for training and the official test set of 49,606 examples. Each sample is annotated with a list of gold entity aliases derived from Wikidata, which serves as the reference set for our verifiable reward. Crucially, **the training and test sets share no overlapping entities**. The dataset covers ten language pairs (English → X), detailed in Table 1. For general translation capability, we evaluated on **WMT24++** (Deutsch et al., 2025) across the corresponding languages, using the official test sets without any domain-specific fine-tuning. Further details on data composition are provided in Appendix E.

Models and Baselines. We employed Qwen3-8B and Qwen3-14B (Yang et al., 2025) as our backbone models, which are pre-trained on 36T tokens and possess native reasoning capabilities. Our proposed EA-RLVR was compared against two primary internal baselines: (1) the base model and (2) Supervised Fine-Tuning (SFT) on the same 7k examples to control for data exposure. To further contextualize our performance, we included several strong external baselines: (i) frontier proprietary LLMs: GPT-5-mini; (ii) strong open-source model: Qwen3-235B-A22B, Marco-o1 (Zhao et al., 2024a), a multilingual reasoning model, and DeepTrans-7B (Wang et al., 2025a), a specialized reasoning-based translation model. More evaluation details are in Appendix E.

Training and Implementation. We implemented EA-RLVR using the `verl` framework (Sheng et al., 2024). All EA-RLVR models were trained using the policy optimization algorithm described in §3.3. We applied full-parameter tuning across all our experiments, including both the EA-RLVR and the baselines. Full implementation details are provided in Appendix E.

Evaluation Metrics. We report three primary metrics: (1) **Entity Translation Accuracy**: Consistent with the normalized substring matching defined in Eq. 1, Entity Translation Accuracy measures the percentage of test samples where the generated translation y successfully includes the

Table 1: Entity translation accuracy (%) on XC-Translate across ten language directions (en → X). Train and test sets share no overlapping entities. RLVR consistently outperforms SFT under the same data budget, while also maintaining strong character-level faithfulness.

Model	Entity Translation Accuracy on XC-Translate (*en* → *X*)										**Avg.**
	ar	de	es	fr	it	ja	ko	th	tr	zh	**Acc/chrF**
Baselines											
GPT-5-mini	35.03	36.03	42.71	35.85	35.37	37.27	30.20	15.96	40.44	29.85	33.87/62.30
Qwen3-235B-A22B	27.93	32.06	41.79	34.29	33.07	29.07	30.22	15.32	34.32	32.38	31.05/61.99
Marco-o1	11.88	19.72	26.47	22.09	21.26	11.77	8.42	3.45	17.71	16.23	15.90/49.68
DeepTrans-7B	11.41	21.58	30.10	23.49	22.44	13.21	8.44	2.90	17.37	15.96	16.69/47.88
Ours											
Qwen3-8B	14.91	23.64	31.04	25.97	25.01	17.13	11.12	5.63	23.94	23.50	20.19/56.07
+ SFT	15.28	23.28	30.76	25.32	25.44	16.86	11.12	5.86	23.56	23.08	20.06/56.20
+ **EA-RLVR**	**25.23**	**31.01**	**44.44**	**34.89**	**35.35**	**24.02**	**25.70**	**14.74**	**27.79**	**29.31**	**29.25/59.86**
Qwen3-14B	20.48	26.86	34.88	28.66	28.70	18.81	17.67	7.98	25.96	26.57	23.66/58.22
+ SFT	20.21	26.97	35.13	28.75	30.05	18.32	17.65	8.01	25.93	27.75	23.88/59.43
+ **EA-RLVR**	**28.17**	**33.71**	**44.70**	**35.10**	**37.62**	**27.64**	**31.29**	**17.09**	**29.96**	**33.42**	**31.87/62.27**

correct cultural entity (i.e., $m(y, \mathcal{G}(x)) = 1$). (2) **chrF**: We report the sentence-level Character F-score (chrF) (Popović, 2015) to assess the overall quality of the generated translations. (3) **XCOMET-XL** (Guerreiro et al., 2024): A state-of-the-art reference-based neural metric used to assess the general quality and fluency of the translations on WMT24++.

4.2 EA-RLVR incentivizes parametric knowledge

Table 1 reports the entity translation accuracy on the XC-Translate test set. The results provide empirical support for our hypothesis regarding parametric knowledge activation, revealing a fundamental divergence in effectiveness between imitation-based and reasoning-based optimization.

Breaking the Imitation Ceiling via Reasoning. Standard SFT yields negligible gains (e.g., +0.22% for Qwen3-14B), despite using the same 7k data as EA-RLVR. This outcome is predictable as our train-test entities are disjoint. Unlike style transfer, where learning generalizable patterns suffices, entity translation inherently biases towards memorization, limiting the effectiveness of SFT. EA-RLVR, however, reframes this task as a reasoning problem. Consequently, it achieves substantial improvements (+8.21%), **enabling the 14B model (31.87%) to outperform the much larger Qwen3-235B-A22B baseline (31.05%).** RLVR never presents the gold entity to the model during training, therefore its performance gains cannot stem from memorizing entity mappings.

Table 2: Effect of task-specific prompting vs. EA-RLVR on entity translation accuracy (%, Qwen3-8B).

Configuration	**Avg. ETA**	**Δ**
Standard Prompt	20.19	–
Task-Specific Prompt	21.94	+1.75
Standard Prompt + EA-RLVR	**29.25**	**+9.06**

Beyond Prompt Engineering. Since Qwen3 natively supports chain-of-thought reasoning, one might ask if similar gains could be obtained simply through better prompting. To test this, we evaluated Qwen3-8B with a task-specific prompt that explicitly instructs the model to identify cultural entities, deliberate on their translations, and then translate (Table 2). While such guided prompting yields a modest improvement (+1.75%), it falls far short of EA-RLVR (+9.06%), which uses the same standard translation prompt as the base model. This confirms that the gains stem from internalized reasoning strategies acquired through RL, not from surface-level instruction following.

4.3 Transfer to General Translation

A potential concern with specialized reinforcement learning is the risk of "alignment tax," where optimizing for a narrow objective (entity correctness) degrades general capabilities. We investigate this on the WMT24++ benchmark (Table 3) and observe the opposite effect.

Reasoning Improves General Quality. Despite being trained solely on the 7k XC-Translate train set, EA-RLVR models consistently improve general translation quality across all evaluated lan-

Table 3: XCOMET-XL score on WMT24++ across ten language directions. All models are trained only on the XC-Translate cross-cultural entity dataset, without using WMT data or general MT supervision. Despite this, EA-RLVR consistently improves performance on general machine translation, and further benefits appear when scaling to the full XC-Translate dataset.

Model	XCOMET score on WMT24++ ($en \rightarrow X$)										**Avg.**
	ar	de	es	fr	it	ja	ko	th	tr	zh	
Baselines											
GPT-5-mini	72.86	91.71	86.93	84.30	86.34	82.15	82.56	80.29	79.49	77.76	82.44
Qwen3-235B-A22B	69.72	90.42	85.53	81.93	84.59	78.54	79.76	77.32	73.92	75.88	79.76
Marco-o1	53.50	84.09	80.23	75.38	74.84	67.04	67.04	67.04	48.15	70.47	68.78
DeepTrans-7B	56.51	84.69	81.45	76.31	74.86	70.24	62.94	65.80	48.78	71.53	69.31
Ours											
Qwen3-8B	61.88	88.06	82.36	78.33	81.09	72.86	71.97	72.43	62.48	73.11	74.46
+ SFT	62.73	88.29	82.91	78.52	80.23	72.99	71.73	72.20	62.47	73.67	74.57
+ EA-RLVR	**64.65**	**88.85**	**83.29**	**79.48**	**81.16**	**73.60**	**74.04**	**72.92**	**64.91**	**74.20**	**75.71**
+ SFT (full data)	62.51	87.78	82.25	78.79	80.48	72.32	71.37	71.57	61.83	73.31	74.22
+ EA-RLVR (full data)	**65.29**	**89.00**	**83.48**	**79.82**	**82.17**	**74.52**	**73.62**	**73.88**	**65.58**	**74.06**	**76.14**
Qwen3-14B	66.37	89.29	83.79	80.15	81.91	76.02	75.73	74.87	67.40	75.10	77.06
+ SFT	66.20	89.12	84.10	80.09	82.37	76.38	75.35	75.23	67.70	75.15	77.17
+ EA-RLVR	**68.30**	**89.79**	**84.28**	**80.78**	**83.56**	**77.61**	**77.49**	**76.30**	**70.16**	**75.84**	**78.41**
+ SFT (full data)	65.27	89.11	84.07	79.87	82.47	76.26	75.50	74.92	67.82	75.18	77.05
+ EA-RLVR (full data)	**68.00**	**90.11**	**85.06**	**81.27**	**84.10**	**78.14**	**77.57**	**75.97**	**70.25**	**75.99**	**78.65**

guages. Qwen3-14B + EA-RLVR achieves an average XCOMET score of 78.41, a +1.35 point improvement over the base model. This suggests that the reasoning strategies learned for entity translation—such as attending more carefully to source semantics and deliberating before generating—are transferable. The model becomes less prone to literal translation and more faithful to the source text, benefiting general translation tasks.

Scalability and Robustness. Table 3 also compares models trained on the standard 7k set versus the full dataset. While SFT performance stagnates or even slightly degrades when scaling data (likely due to overfitting on the specific formatting of the entity dataset), EA-RLVR continues to improve. The "full data" setting yields further gains, pushing the average XCOMET score to 78.65 for the 14B model. This indicates that our outcome-based reward formulation provides a stable optimization landscape that scales effectively with data, unlike SFT which may suffer from distribution shift.

5 Analysis

5.1 Improved Sampling Efficiency: Unlocking Dormant Knowledge

To analyze the mechanism behind the performance gains, we adopt the $pass@k$ evaluation framework recently utilized to study the boundaries of RLVR in reasoning tasks (Yue et al., 2025). Formally, $pass@k$ estimates the probability that at least one correct translation exists within k independent samples generated for a given input. Following (Chen et al., 2021), we calculate the unbiased estimator (see Appendix F for details). By observing how this probability scales with k, we can distinguish between *knowledge injection* (learning new information) and *knowledge activation* (surfacing existing information). Figure 3 compares the entity translation accuracy of Qwen3-8B (Base) and EA-RLVR across varying sample sizes $k \in [1, 128]$.

We observe a distinct convergence pattern across most languages. At $k = 1$, EA-RLVR holds a substantial lead over the base model, confirming that our policy optimization successfully concentrates probability mass on the correct entity translations. However, as k increases, the base model's accuracy rises steeply, often converging with or appearing to surpass the RLVR model at $k = 128$. This phenomenon has a critical implication: **the base model inherently possesses the necessary cultural knowledge** to translate these entities correctly (evidenced by high performance at large k), but it fails to rank these correct translations as the most probable candidates during standard decoding. EA-RLVR functions as a steering mechanism that activates this dormant knowledge, transforming low-probability correct candidates into high-probability deterministic outputs, rather than memorizing new mappings from external supervision. We further isolate the contribution of the explicit reasoning phase in Appendix B, finding that the

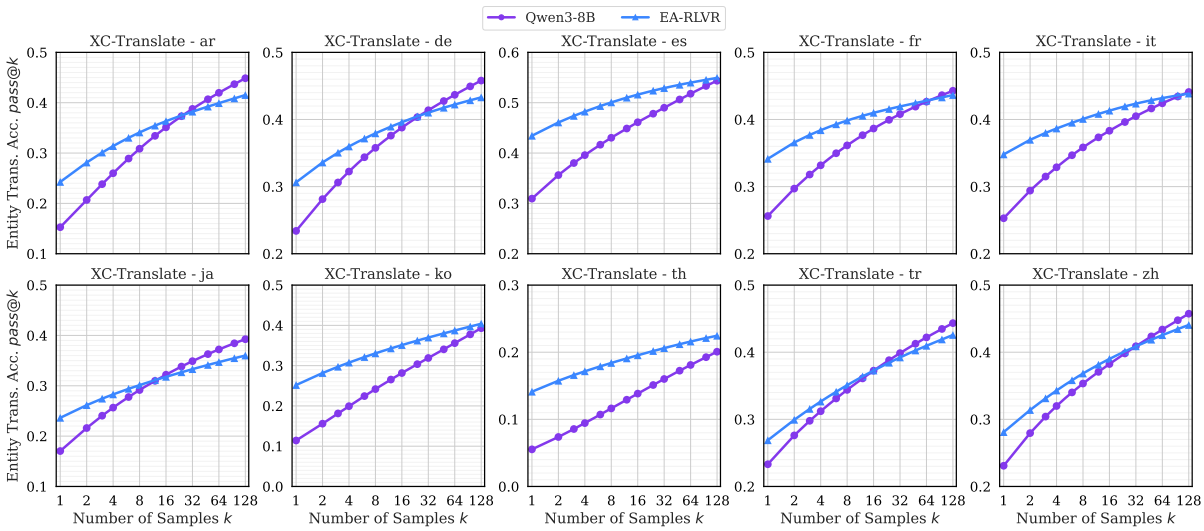

Figure 3: Entity Translation Accuracy $pass@k$ curves across ten languages. The Base model (purple) shows poor performance at $k = 1$ but improves rapidly as k increases, indicating latent knowledge is present but buried. EA-RLVR (blue) significantly boosts $pass@1$ accuracy, effectively surfacing this parametric knowledge. The convergence at high k confirms that improvements stem from better utilization of existing knowledge rather than learning new facts.

"thinking" workspace is essential for absorbing the complexity of cultural alignment without sacrificing general fluency.

Sampling Efficiency and Determinism. The slope of the curves in Figure 3 further elucidates the shift in model behavior. The base model exhibits a high-entropy distribution over entities, requiring extensive sampling ($k \gg 1$) to uncover the correct answer. In contrast, the RLVR curves are notably flatter, indicating a more deterministic policy where the model is confident in its reasoning path. While high determinism theoretically reduces diversity (explaining the slight underperformance at $k = 128$ in some high-resource languages where the base model's broad search space is advantageous), it is the desired behavior for a translation system: users expect the correct cultural translation in a single attempt ($k = 1$), not after filtering through a hundred generations. EA-RLVR effectively optimizes for this *sampling efficiency*.

5.2 The Fluency Trap: Neural Rewards Fail Cultural Entities

A natural alternative to our rule-based framework is to optimize state-of-the-art neural quality metrics directly. To investigate this, we trained a **Comet-RL** baseline on Qwen3-8B using the same RL setup but replacing our normalized entity matching reward with sentence-level comet scores, specifically the wmt22-comet-da (Rei et al., 2022). Figure 4 presents a striking divergence in optimization outcomes, revealing what we term the *"Fluency Trap"*:

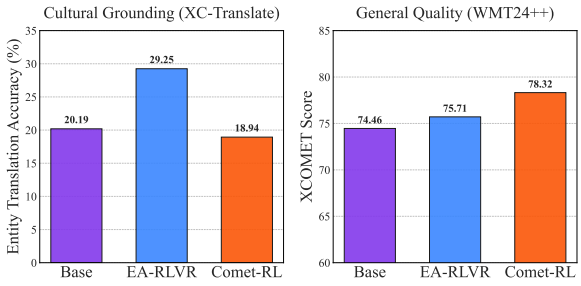

Figure 4: **Verifiable vs. Neural Rewards.** While Comet-RL maximizes general fluency (Right) at the cost of cultural accuracy (Left), falling into a "fluency trap", EA-RLVR achieves robust improvements across both cultural grounding and general translation quality.

High Fluency, Low Grounding. As shown in the right panel, the Comet-RL model achieves substantial gains in general translation quality, boosting the XCOMET score on WMT24++ from 74.46 to 78.32. This confirms that RL effectively optimized the reward signal. However, the left panel reveals a critical failure: on the entity-dense XC-Translate benchmark, Comet-RL's entity accuracy actually *degrades* from 20.19% (Base) to 18.94%. The neural metric, lacking fine-grained resolution for specific entities, fails to penalize these fluent but culturally incorrect entity errors, which echos

prior observations (Rei et al., 2023).

Verifiable Rewards Ensure Alignment. In contrast, EA-RLVR escapes this trap. By anchoring supervision on verifiable outcomes, it forces the model to prioritize semantic correctness. This yields a massive improvement in entity accuracy (+9.06% absolute) while still conferring robust gains in general translation quality (+1.25 XCOMET). This result fundamentally justifies our approach: while holistic neural metrics drive fluency, verifiable constraints are indispensable for aligning models with entity translation tasks.

5.3 Cross-Lingual Generalization

The preceding analyses establish that EA-RLVR activates dormant parametric knowledge (§5.1) through verifiable rewards (§5.2). If the acquired strategy is indeed a generalizable reasoning skill rather than a set of language-specific entity mappings, it should transfer across typologically distinct language families. To test this prediction, we partition the ten target languages into two groups: Group A (Asian & Semitic: ar, ja, ko, th, zh) and Group B (European & Turkic: de, es, fr, it, tr), train on one group exclusively, and evaluate on the other without any target-language supervision.

As shown in Table 4, we observe robust positive transfer in both directions: entity accuracy improves by +7.27% (A→B) and +6.28% (B→A) over the base model, despite the two groups sharing neither scripts nor typological features. This indicates that EA-RLVR does not memorize language-specific entity mappings but instead induces a generalizable reasoning strategy: grounding cultural contexts within the thinking trace before generating the translation (Appendix G). Such cross-lingual transfer of *reasoning ability* distinguishes our approach from conventional multilingual MT, where improvements are typically confined to the supervised language pairs.

6 Conclusion

In this work, we investigate the underutilized potential of parametric knowledge in cross-cultural translation: while LLMs possess extensive latent cultural knowledge ($pass@128$ performance), they struggle to utilize it during standard decoding ($pass@1$). Motivated by this observation, we propose **EA-RLVR**, a framework that transforms cross-cultural translation from a memorizing task into a reasoning-intensive process. Our extensive experiments reveals that incentivizing verifiable correctness is superior to optimizing neural quality metrics, which often lead models into a "fluency trap", generating smooth but culturally inaccurate entities. By anchoring supervision on entity matching and allowing the model to reasoning, EA-RLVR enables a 14B model to outperform a 235B baseline significantly on unseen entities. Ultimately, our findings suggest a potential paradigm shift for knowledge-intensive translation: moving beyond mere imitation of references (SFT) or reliance on external retrieval, toward internalizing self-evolving strategies that effectively unlock the model's inherent potential.

Acknowledgment

The present research was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0203000). We would like to thank the anonymous reviewers for their insightful comments.

Limitations

Gap to Latent Potential. While EA-RLVR significantly outperforms SFT, a notable disparity remains between the optimized policy's expected single-sample performance ($pass@1$) and the model's theoretical upper bound estimated by rejection sampling ($pass@128$, as shown in Figure 1). This gap highlights a shared limitation across current RLVR methodologies: standard policy gradient algorithms often struggle to fully explore and converge to the global optimum within a limited sample budget. Future research could bridge this gap by developing more sample-efficient optimization algorithms or by scaling the number of rollout trajectories (G). Orthogonally, knowledge-enriched pre-training strategies that co-organize semantically related documents (Zhou et al., 2026) have been shown to improve $pass@k$ by strengthening the parametric knowledge base itself; combining such pre-training with EA-RLVR's policy-level activation is a promising direction. Although computationally intensive, expanding the exploration horizon offers a promising avenue for approaching the model's intrinsic capability ceiling (Pan et al., 2025b).

Knowledge Boundaries. Our framework is designed for *knowledge elicitation*, not *knowledge injection*. EA-RLVR optimizes the retrieval of long-tail cultural concepts that exist within the model's

Table 4: Cross-lingual generalization on XC-Translate with Qwen3-8B. The model is trained on one group of languages and evaluated on both. **Bold text** indicates zero-shot cross-lingual transfer (e.g., Train A → Test B), while gray text indicates in-domain performance. Improvements over the Base model on unseen language groups demonstrate the acquisition of a transferable reasoning strategy.

Train split	*Group A*					*Group B*					**Group A**	**Group B**
(Qwen3-8B)	ar	ja	ko	th	zh	de	es	fr	it	tr	**Avg.**	**Avg.**
Base	14.91	17.13	11.12	5.63	23.50	23.64	31.04	25.97	25.01	23.94	14.46	25.92
Group A	48.27	43.48	47.70	25.76	53.24	**29.70**	**40.61**	**33.78**	**34.95**	**26.89**	43.69	**33.19**
Group B	**24.81**	**21.77**	**18.32**	**8.62**	**30.20**	51.28	62.50	49.66	51.61	50.19	**20.74**	53.05

pre-training data but are suppressed during standard decoding. Consequently, it cannot generate correct translations for entities entirely absent from the pre-training corpus. However, we observed that our method synergizes effectively with retrieval-augmented systems rather than acting as a simple alternative, providing a combined benefit that we evaluate in Appendix C.

Reward Rigidity vs. Flexibility. We prioritize optimization stability via rigid substring matching to prevent the reward hacking often observed with neural metrics. Although we mitigate the risk of false negatives by employing a comprehensive gold set of aliases (derived from Wikidata) rather than a single reference, this strict verification process may still occasionally penalize valid but unlisted stylistic variations. Developing rewards that balance verifiable strictness with semantic flexibility remains an open challenge for the field.

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, Toronto, Canada. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360.

Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, and 1 others. 2025. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Preprint*, arXiv:2205.14135.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2026. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, and 1 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284.

Tianyu Dong, Bo Li, Jinsong Liu, Shaolin Zhu, and Deyi Xiong. 2025. Mlas-lora: language-aware parameters detection and lora-based knowledge transfer for multilingual machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15645–15660.

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

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

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.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2022. DEEP: DEnoising entity pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1753–1766, Dublin, Ireland. Association for Computational Linguistics.

Fanding Huang, Guanbo Huang, Xiao Fan, Yi He, Xiao Liang, Xiao Chen, Qinting Jiang, Faisal Nadeem Khan, Jingyan Jiang, and Zhi Wang. 2026. Semantic-space exploration and exploitation in rlvr for llm reasoning. *Preprint*, arXiv:2509.23808.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Chenyang Si, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, and 1 others. 2025a. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. *arXiv preprint arXiv:2506.00979*.

Changjiang Jiang, Xinkuan Sha, Fengchang Yu, Jingjing Liu, Jian Liu, Mingqi Fang, Chenfeng Zhang, and Wei Lu. 2026. Fake-hr1: Rethinking reasoning of vision language model for synthetic image detection. In *ICASSP*.

Changjiang Jiang, Fengchang Yu, Haihua Chen, Wei Lu, and Jin Zeng. 2025b. Tabdsr: Decompose, sanitize, and reason for complex numerical reasoning in tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3172–3196.

Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. 2025. Revisiting entropy in reinforcement learning for large reasoning models. *arXiv preprint arXiv:2511.05993*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Tian Liang, Xing Wang, Mingming Yang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Addressing entity translation problem via translation difficulty and context diversity. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11628–11638, Bangkok, Thailand. Association for Computational Linguistics.

Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, and Yang Liu. 2025. DAPO : Improving multi-step reasoning abilities of large language models with direct advantage-based policy optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Leiyu Pan, Bojian Xiong, Lei Yang, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, and 1 others. 2025a. Advancing large language models for tibetan with curated data and continual pre-training. *arXiv preprint arXiv:2507.09205*.

Leiyu Pan, Bojian Xiong, Lei Yang, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, and 1 others. 2025b. Banzhida: Advancing large language models for tibetan with curated data and continual pre-training. *arXiv e-prints*, pages arXiv–2507.

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.

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

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. Association for Computational Linguistics.

Matiss Rikters and Makoto Miwa. 2024. Entity-aware multi-task training helps rare word machine translation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 47–54,

Tokyo, Japan. Association for Computational Linguistics.

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

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. Drt: Deep reasoning translation via long chain-of-thought. *Preprint*, arXiv:2412.17498.

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

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

Xinwei Wu, Heng Liu, Xiaohu Zhao, Yuqi Ren, Linlong Xu, Longyue Wang, Deyi Xiong, Weihua Luo, and Kaifu Zhang. 2026. Finding the translation switch: Discovering and exploiting the task-initiation features in llms. *arXiv preprint arXiv:2601.11019*.

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*.

Lei Yang, Wei Bi, Chenxi Sun, Renren Jin, and Deyi Xiong. 2026. SOUP: Token-level Single-sample Mix-policy Reinforcement Learning for large language models. *CoRR*, abs/2601.21476.

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

Linhao Yu, Tianmeng Yang, Siyu Ding, Renren Jin, Naibin Gu, Xiangzhao Hao, Shuaiyi Nie, Deyi Xiong, Weichong Yin, Yu Sun, and 1 others. 2026. Knowrl: Boosting llm reasoning via reinforcement learning with minimal-sufficient knowledge guidance. *arXiv preprint arXiv:2604.12627*.

Xinbin Yuan, Jian Zhang, Kaixin Li, Zhuoxuan Cai, Lujian Yao, Jie Chen, Enguang Wang, Qibin Hou, Jinwei Chen, Peng-Tao Jiang, and Bo Li. 2025. SEGUI: Enhancing visual grounding for GUI agents via self-evolutionary reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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? *Preprint*, arXiv:2504.13837.

Qiannian Zhao, Chen Yang, Jinhao Jing, Yunke Zhang, Xuhui Ren, Lu Yu, Shijie Zhang, and Hongzhi Yin. 2026. Know what you know: Metacognitive entropy calibration for verifiable rl reasoning. *arXiv preprint arXiv:2602.22751*.

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

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024b. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Jiang Zhou, Yunhao Wang, Xing Wu, Tinghao Yu, and Feng Zhang. 2026. Wrap++: Web discovery amplified pretraining. *arXiv preprint arXiv:2604.06829*.

A Impact of Structural Gates

To validate the necessity of the structural constraints introduced in Section 3.2, we conduct an ablation study using the full XC-Translate dataset with extended optimization (1,000 steps). We isolate the contributions of the format and length gates by comparing three configurations:

- **Soft Format**: A relaxed baseline where any output containing a valid `<think>` block is eligible for rewards, with no length penalty applied.
- **Format Gate Only**: The strict format verification (g_{fmt}) is applied, but the relative length constraint (g_{len}) is removed.
- **EA-RLVR**: Our proposed framework, which enforces both strict format verification and the relative length constraint (g_{len}).

Training Dynamics and Stability. The training dynamics, visualized in Figure 5, demonstrate that structural constraints are critical for optimization stability. The most prominent difference lies in the *Response Length* (Center). In the absence of the length constraint (g_{len}), both ablated variants suffer from a catastrophic "length explosion," where generation length increases uncontrollably. This instability is mirrored in the *Actor Entropy* (Right), where the ablated models exhibit erratic spikes, indicating that the policy fails to converge to a stable reasoning strategy. In contrast, EA-RLVR maintains a consistent length and stable entropy profile throughout training.

Reward Hacking via Enumeration. Qualitative analysis reveals that the length explosion is a symptom of reward hacking. As illustrated in the case study (Figure 6), without the length penalty, the optimization landscape encourages a degenerate solution: the model learns to "brute-force" the verification condition ($m(y, \mathcal{G}(x))$) by enumerating synonymous entities or repeating candidates. This strategy maximizes the recall of the gold entity at the expense of precision and structural integrity. By treating the translation task as a keyword-stuffing exercise, the model achieves technically high rewards but produces unusable translations.

The Deception of High Rewards. The *Average Reward* curves (Left) present a counter-intuitive trend: the weaker constraints yield higher raw reward values. The *Soft Format* setting achieves the highest reward trajectory despite exhibiting the earliest collapse in generation quality. This phenomenon is a classic manifestation of **Goodhart's Law**: when the unconstrained entity-match metric becomes the sole target, the model exploits its loopholes (e.g., infinite generation) rather than improving the intended task utility. EA-RLVR's lower reward curve reflects a constrained, harder-to-optimize landscape that successfully steers the model away from these degenerate local optima and toward concise, correct translations.

B Impact of Reasoning: Does Thinking Matter?

A central premise of EA-RLVR is that a dedicated reasoning phase ("thinking") allows the model to navigate the optimization landscape more effectively than immediate generation. To isolate the impact of this reasoning process, we conduct a controlled comparison using the Qwen3-4B-2507 family.

Experimental Control. We compare two specific checkpoints: the standard instruction-tuned model, `Qwen3-4B-Instruct-2507`, and the reasoning-enhanced model, `Qwen3-4B-Thinking-2507`. We apply the EA-RLVR framework to both models with a crucial adaptation for the Instruct baseline: since `Qwen3-4B-Instruct-2507` does not generate reasoning traces (i.e., no `<think>` tokens), we remove the format verification gate (g_{fmt}) from the reward function. However, to ensure a fair comparison of supervision signals, we retain both the length constraint (g_{len}) and the entity matching reward ($m(y, \mathcal{G}(x))$). This setup allows us to strictly evaluate whether the presence of a thinking process facilitates better alignment with the verifiable reward.

Thinking Unlocks Higher Entity Accuracy. Figure 7 (Center) illustrates the progression of entity translation accuracy on the XC-Translate test set. The *Non-Think Model* (Cyan) plateaus early at approximately 21% accuracy. In contrast, the *Think Model* (Blue) achieves a significantly higher peak of $\sim$26%, despite starting from a lower baseline. This confirms that the reasoning trace provides the necessary computational workspace to resolve complex cultural entity mappings that are inaccessible to a single-pass decoding policy.

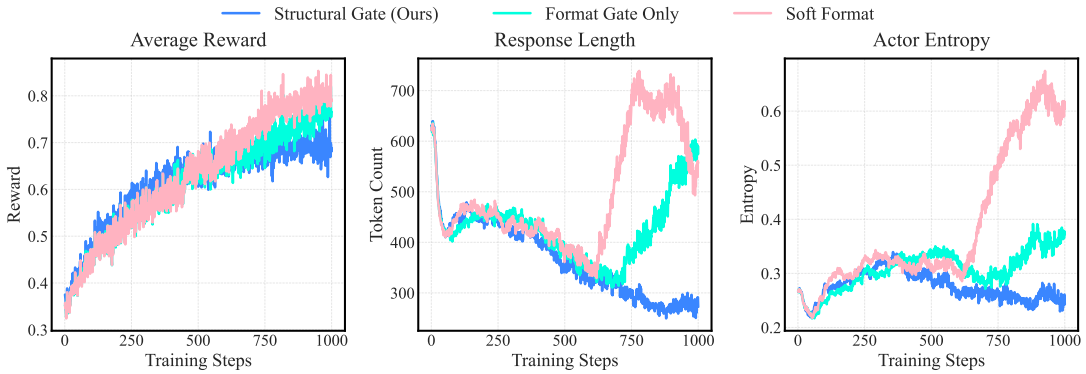

Figure 5: **Ablation Dynamics.** Training curves for Average Reward, Response Length, and Actor Entropy. Without the full structural gates (EA-RLVR, Blue), the model suffers from reward hacking, characterized by an explosion in response length (Center) and unstable entropy (Right), despite achieving higher raw rewards (Left).

Dynamics of Length and Stability. The training dynamics reveal a crucial interaction between reasoning length and reward. As shown in Figure 7 (Left), the Think Model initially receives near-zero rewards. This is an artifact of the base model's instability: the untrained `Qwen3-4B-Thinking-2507` frequently generates extremely long chain-of-thought traces that exceed our training context window of 4096 tokens, causing the samples to be truncated and penalized. However, EA-RLVR rapidly corrects this behavior. The *Response Length* curve (Right) shows a dramatic reduction in token count within the first 50 steps. The model learns to be concise, condensing its reasoning into an efficient path that fits the constraints while maximizing the entity-match reward. This demonstrates that EA-RLVR serves not only as a task optimizer but also as a length regularizer for reasoning models.

Thinking Mitigates the "Alignment Tax". We further evaluate the impact of these strategies on general translation quality using WMT24++. Table 5 presents the XCOMET scores. A striking divergence is observed:

- **Standard Model (Instruct):** Applying EA-RLVR to `Qwen3-4B-Instruct-2507` leads to a slight regression in general quality (Avg. 89.11 $\rightarrow$ 88.46). Without a reasoning buffer, the model is forced to overload its generation weights to satisfy the strict entity constraints, leading to a "fluency tax" where general translation quality is sacrificed.
- **Reasoning Model (Thinking):** Conversely, `Qwen3-4B-Thinking-2507` *improves* with EA-RLVR (Avg. 90.36 $\rightarrow$ 90.67). The reasoning trace absorbs the complexity of the entity task, allowing the final translation generation to remain fluent and robust.

Note on Evaluation Subset. It is important to note that the baseline `Qwen3-4B-Thinking-2507` is highly unstable for translation tasks, often generating endless thought loops exceeding 32k tokens. Consequently, it fails to produce any translation for a large portion of the WMT24++ test set. To ensure a scientifically valid comparison, the results in Table 5 are calculated on the **common intersection** of sentences where the baseline model successfully produced an output. Table 6 details the data statistics. The valid subset size ranges from 68 to 260 samples per language (out of 960), highlighting the severity of the length explosion issue in the baseline model and the necessity of this filtering step.

C Synergy with Retrieval-Augmented Generation

A central question in modern translation systems is the interplay between optimizing internal parameters (via RLVR) and utilizing external non-parametric knowledge (via RAG). To determine whether our method complements retrieval-based approaches, we conduct an ablation study using a standard RAG pipeline.

Experimental Setup. We employ **mContriever** (Izacard et al., 2021), a widely used multilingual dense retriever, without any task-specific fine-tuning. We index the aliases of entities present in the XC-Translate test set (extracted from Wikidata via the provided QIDs). For each source sentence, we retrieve the top-k most relevant entity aliases and prepend them to the system prompt as

Table 5: Impact of reasoning on general translation quality (XCOMET on WMT24++). Due to the instability of the baseline Qwen3-4B-Thinking model (which frequently generates infinite reasoning traces exceeding 32k tokens), this evaluation is restricted to the **common subset** of sentences where the baseline model successfully produced a valid output. On this subset, EA-RLVR improves the reasoning model's quality, whereas it degrades the standard model, suggesting that a thinking workspace is necessary to absorb the complexity of entity constraints without sacrificing fluency.

Model	**XCOMET score on WMT24++ Subset** ($en \rightarrow X$)										**Avg.**
	ar	de	es	fr	it	ja	ko	th	tr	zh	
Standard Instruction Backbone											
Qwen3-4B-Instruct	87.09	95.14	92.01	89.75	90.78	89.70	86.69	90.37	86.25	83.32	89.11
+ **EA-RLVR**	87.41	95.17	91.12	88.21	89.79	88.37	84.65	90.03	86.94	82.90	88.46
Reasoning Backbone											
Qwen3-4B-Thinking	88.84	95.74	92.48	90.20	92.21	90.66	**89.39**	**91.04**	**90.38**	82.65	90.36
+ **EA-RLVR**	**89.03**	**96.17**	**93.25**	**91.16**	**93.09**	**90.72**	88.66	90.79	89.69	**84.19**	**90.67**

Table 6: Statistics of the evaluation subset for WMT24++. Due to the endless thinking issue, the `Qwen3-4B-Thinking` baseline yields valid outputs for only a fraction of the test set. We report results on the **Common** intersection to ensure fair comparison.

Language Pair	**Total (Test Set)**	**Common Subset**
en-ar	960	68
en-de	960	143
en-es	960	260
en-fr	960	193
en-it	960	211
en-ja	960	144
en-ko	960	135
en-th	960	124
en-tr	960	99
en-zh	960	241

context. For the combined setting (**EA-RLVR + RAG**), we utilize the Qwen3-8B model trained via EA-RLVR and provide it with the same retrieved context during inference.

Parametric Optimization Outperforms Naive Retrieval. As shown in Table 7, the standard RAG baseline improves the base model's entity accuracy from 20.19% to 23.14% and slightly boosts general quality (chrF 57.43), validating the effectiveness of our retrieval setup. However, **EA-RLVR alone significantly outperforms the RAG baseline** (29.25%). This result highlights a critical insight: for cross-cultural translation, the bottleneck is often not the *availability* of knowledge (which RAG provides), but the model's ability to *align* that knowledge with the translation context. EA-RLVR addresses this alignment directly via optimization, proving more effective than passively injecting context.

Additive Gains. Crucially, the two approaches are synergistic. The **EA-RLVR + RAG** configuration achieves the highest overall performance (30.49% Acc, 60.05 chrF). This demonstrates that the reasoning patterns learned by EA-RLVR are robust; the model does not "overfit" to its internal weights but retains the flexibility to incorporate external evidence. By transforming the model into an active reasoner, EA-RLVR enables it to utilize retrieved context to resolve tail cases that neither parametric knowledge nor retrieval could solve alone.

D Broad Knowledge Activation Across Cultural Categories

To verify that the gains are not confined to a narrow domain, Table 8 breaks down entity translation accuracy by the 14 cultural categories annotated in XC-Translate. EA-RLVR yields consistent improvements across *all* categories, from musical works (+19.84%) and natural places (+19.23%) to book series (+1.76%). This breadth confirms that the activated parametric knowledge spans diverse cultural domains rather than reflecting a bias toward any single entity type.

Further Discussion Our finding that verifiable rewards activate dormant parametric knowledge and induce transferable reasoning strategies resonates with concurrent advances across several domains. In the context of LLM reasoning, Zhao et al. (2026) show that metacognitive entropy calibration substantially improves verifiable RL reasoning by helping models better distinguish what they know from what they do not, aligning with our observation that explicit verification signals help surface latent knowledge. Similarly, Yu et al.

Table 7: **Synergy between Parametric and Non-Parametric Knowledge.** Comparison of EA-RLVR against a standard Retrieval-Augmented Generation (RAG) baseline on Qwen3-8B. Adding RAG to the base model yields moderate gains, validating the retrieval setup. However, EA-RLVR acting as a standalone method provides a substantially larger improvement. Crucially, the combined setting (**EA-RLVR + RAG**) achieves the highest accuracy and faithfulness (chrF), indicating that the reasoning capabilities induced by EA-RLVR effectively complement external knowledge retrieval.

Model	Entity Translation Accuracy on XC-Translate ($en \rightarrow X$)										**Avg.**
	ar	de	es	fr	it	ja	ko	th	tr	zh	**Acc/ChrF**
Qwen3-8B	14.91	23.64	31.04	25.97	25.01	17.13	11.12	5.63	23.94	23.50	20.19/56.07
+ RAG	17.13	32.85	36.62	30.58	28.82	16.35	11.79	6.18	27.81	23.25	23.14/57.43
+ EA-RLVR	25.23	31.01	44.44	34.89	35.35	24.02	25.70	14.74	27.79	29.31	29.25/59.86
+ EA-RLVR + RAG	**26.17**	**35.50**	**47.81**	**36.63**	**37.43**	**23.36**	**25.52**	**14.42**	**29.33**	**28.75**	**30.49/60.05**

Case: Reward Hacking

User: Translate the follow text to Chinese (Traditional): Sunday Cont.
LLM: <think> Okay, "Sunday Cont." in Chinese is "星期日連續劇" or "周日連續劇"?
<\think>

週日連續劇 (或 "週日集數" 或 "周日集" 也可以是 "主日集" 或 "星期日賽事", 但常見的可能是 "主日連續劇" 或 "週日特別節目" 或 "週日連續劇"。)

But the common term might be "週日劇集" or "周日連續劇" or "主日連續劇" or "Sunday Show" or "星期日連續劇".

The most common translation in Chinese is "週日連續劇" or "週日劇集".

Alternatively, "Sunday Cont." could be "星期日後續" or "週日持續" or "週日繼續".

But since the original text is "Sunday Cont.", it's likely "週日續集" or "主日續集" or "周日續集".

But the best translation would be "週日連續劇" or "週日劇集".

週日集數 (或 "週日節目") or ... ...

[---ENDLESS GENERATION---]

Figure 6: **Reward Hacking Case Study.** In the absence of a length constraint (g_{len}), the model exploits the unconstrained reward by endlessly enumerating possible entity translations to ensure verification success.

(2026) demonstrate that minimal-sufficient knowledge guidance can boost reasoning, complementing our entity-anchored approach where Wikidata aliases serve as the guiding rewards. Beyond text, the RLVR paradigm shows analogous benefits in multimodal settings (Yuan et al., 2025; Jiang et al., 2026, 2025a,b). Together, these results suggest that the core mechanism underlying EA-RLVR, using verifiable rewards to induce reasoning rather than remembering knowledge mapping, may constitute a general principle applicable across modalities. Looking forward, mechanistic interpretability methods such as the translation-switch discovery of Wu et al. (2026) offer a promising lens for uncovering the internal circuits through which EA-RLVR activates parametric knowledge, a direction we leave for future work.

Table 8: Entity Translation Accuracy (%) by cultural category on XC-Translate (Qwen3-14B).

Category	**Base**	**+EA-RLVR**	**Δ**
Musical work	30.15	49.99	+19.84
Natural place	11.54	30.77	+19.23
Plant	36.11	50.93	+14.82
Animal	25.00	38.89	+13.89
Artwork	22.54	32.57	+10.03
Fictional entity	34.16	42.97	+8.81
Person	22.40	30.71	+8.31
Book	24.94	33.00	+8.06
Food	46.35	52.73	+6.38
Landmark	27.78	33.78	+6.00
Movie	12.96	18.85	+5.89
Place of worship	31.86	37.13	+5.27
TV series	12.75	17.62	+4.87
Book series	8.82	10.58	+1.76
MACRO AVG	24.81	34.32	+9.51

E Implementation Details

Data Construction. We conduct experiments on the XC-Translate benchmark across the ten language directions listed in Table 9. As the benchmark does not provide a dedicated training set, we repurpose the official validation set as our training split. Table 10 summarizes the statistics of our data split. Crucially, to strictly evaluate the model's ability to generalize rather than memorize, we ensure that **the training and test sets share no overlapping entities**. The final setup comprises 7,278 examples for training and 49,606 examples for testing.

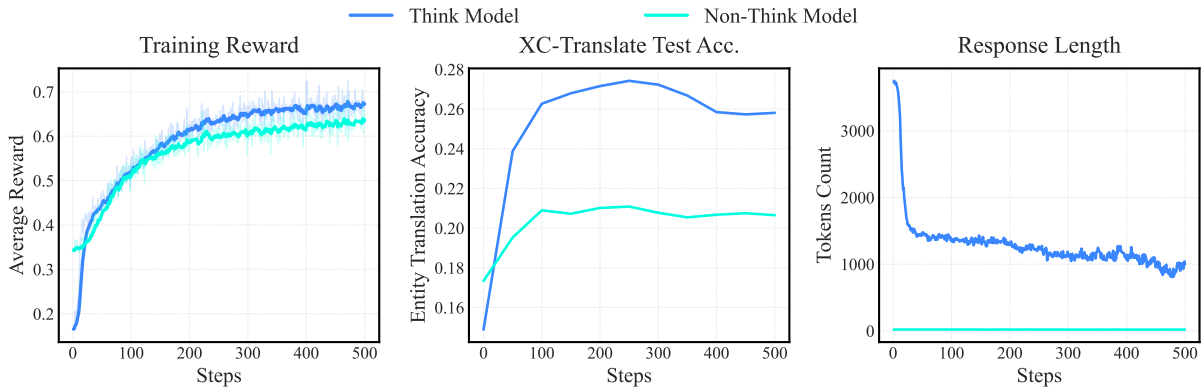

Figure 7: **Training Dynamics of Thinking vs. Non-Thinking Models. (Left)** The Think Model (Blue) initially suffers from low rewards due to context length overflows (> 4096 tokens) but eventually surpasses the Non-Think Model (Cyan). **(Center)** The reasoning capability unlocks a significantly higher ceiling for entity translation accuracy on the XC-Translate test set. **(Right)** EA-RLVR acts as a strong regularizer for reasoning, rapidly curbing the "infinite thought" tendency of the base model to a stable, efficient length.

Table 9: Languages used in our experiments, together with their ISO 639-1 codes, language–region locales, and English names.

ISO 639-1	Locale	English Name
ar	ar_SA	Arabic (Saudi Arabia)
de	de_DE	German (Germany)
es	es_MX	Spanish (Mexico)
fr	fr_FR	French (France)
it	it_IT	Italian (Italy)
ja	ja_JP	Japanese (Japan)
ko	ko_KR	Korean (Korea)
th	th_TH	Thai (Thailand)
tr	tr_TR	Turkish (Turkey)
zh	zh_TW	Chinese (Taiwan, Traditional)

Table 10: Statistics of the dataset used in our experiments. We utilize the official validation set of XC-Translate as our training split. The training and test sets are strictly disjoint in terms of entity coverage.

Language Pair	Train	Test	Total
English → Arabic	722	4,546	5,268
English → Chinese	722	5,181	5,903
English → French	724	5,464	6,188
English → German	731	5,875	6,606
English → Italian	730	5,097	5,827
English → Japanese	723	5,107	5,830
English → Korean	745	5,081	5,826
English → Spanish	739	5,337	6,076
English → Thai	710	3,446	4,156
English → Turkish	732	4,472	5,204
Total	**7,278**	**49,606**	**56,884**

Training Implementation. We implement EA-RLVR using the `verl` library (Sheng et al., 2024), a framework designed for efficient RLHF post-training. Unless otherwise specified, all RL experiments across different model scales (8B, 14B) and variants (Instruct, Thinking) use the unified set of hyperparameters reported in Table 11.

Compute and Environment. All models were trained on 32 × NVIDIA H100 80GB GPUs. The 7k samples training for the 8B model takes approximately 12 hours, while the 14B model takes 24 hours. And the scaled training using full XC-Translate for the 8B model takes approximately 24 hours, while the 14B model takes 48 hours. We use `FlashAttention` for efficient computation (Dao et al., 2022).

SFT Baseline. For the SFT baseline, we fine-tune the base models on the same 7k training examples for 2 epochs, supervised by reference translation. We use a learning rate of 1e-6 with a cosine decay schedule and a global batch size of 64. We SFT our model using ms-swift framework (Zhao et al., 2024b).

EA-RLVR Configuration. Our EA-RLVR optimization follows the critic-free policy gradient approach described in §3.3. We initialize the actor network with the Qwen3 weights post-trained by their official team, which endow the model basic reasoning capability. During the rollout phase, we sample $G = 16$ responses for each prompt to compute the group-normalized advantages.

Table 11 lists the detailed hyperparameters used for the RLVR stage.

Prompt Format. We use the standard chat template of the Qwen3 family. For the input, we

Table 11: Hyperparameters for EA-RLVR training.

Hyperparameter	Value
Optimization	
Optimizer	AdamW
Peak Learning Rate (Actor)	1e-6
Learning Rate Scheduler	Cosine
Warmup Ratio	0.05
Weight Decay	0.1
Train Batch Size	512
PPO Mini Batch Size	128
Total Training Steps	500
PPO / Policy Gradient	
Group Size (G)	16
Clip Ratio ($\varepsilon_{low}, \varepsilon_{high}$)	3e-4,4e-4
Advantage Estimator	Group Normalization
Generation / Rollout	
Sampling Temperature	1.0
Top-p	1.0
Max Sequence Length	4096
Reward Function	
Format Reward (α)	0.2
Length Tolerance (τ)	2.0

wrap the source sentence with the instruction: "`Translate the following sentence into {tgt_lang}, provide only the translated text :...`". For the task-specific prompting ablation (Table 2), we replace the standard instruction with a three-step chain-of-thought prompt: "`Translate the following text from {src_lang} to {tgt_lang}. First, identify any culturally specific entities (such as books, movies, places, or idioms). Second, deliberate on their conventional and culturally appropriate translations in {tgt_lang}. Finally, provide the full translated text:...`"

Evaluation Configuration. We treat the reasoning process (i.e., "thinking") as an intrinsic capability of the models rather than a separate module. Consequently, we enable the thinking mode by default for all applicable models (e.g., Qwen3, Marco-o1 and GPT5-mini) and all settings including SFT and RAG. To ensure a fair comparison, we allocate a unified maximum generation budget of 4,096 tokens for all experiments. Regarding decoding strategies, we follow the best practices established by DeepSeek-R1 (DeepSeek-AI et al., 2026; Yang et al., 2025), setting the sampling temperature to 0.6 and top-p to 0.95. This specific configuration is critical, as lower temperatures (e.g., greedy decoding) tend to induce severe repetition loops and infinite generation behaviors in reasoning-heavy models. Finally, to ensure statistical reliability, all reported results for open-weight models are averaged over three independent runs ($pass@1$).

F Unbiased $pass@k$ Estimator

While $pass@k$ is defined as the probability of generating at least one correct sample in k attempts, directly computing this probability typically requires a very large number of samples to reduce variance. To evaluate $pass@k$ efficiently, we follow the method proposed by Chen et al. (2021). instead of just sampling k times, we generate a larger number of samples n (where $n \geq k$) for each input and count the number of correct samples c. The unbiased estimator for $pass@k$ is then calculated as:

$$\text{pass@}k := \mathbb{E}_{\text{problems}} \left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \right] \tag{7}$$

where $\binom{n}{k}$ denotes the number of combinations of choosing k items from a set of n. Mathematically, this formula calculates the probability that a randomly chosen subset of size k contains at least one correct answer, derived from the complement of the probability that all k chosen samples are incorrect (i.e., chosen from the $n - c$ incorrect samples).

In our experiments, we set the total sample budget $n = 128$ and evaluate $pass@k$ for $k \in \{1, \ldots, 128\}$. If $n - c < k$, the estimator returns 1.0, as it is impossible to choose k incorrect samples.

G Case Study

Qualitative Analysis We present a qualitative analysis of four representative cases to illuminate the mechanism by which EA-RLVR improves entity translation. By examining the generated reasoning traces (denoted as `LLM`), we identify a distinct shift in cognitive patterns: while the baseline model (Qwen3-8B) relies on *literal semantic composition*, EA-RLVR exhibits *entity-aware deliberation* and *domain-specific retrieval*.

Canonicalization of Historical Terminology (Case 1). In Case 1, the user asks for the translation of "Great Ming Code". The baseline Qwen3 adopts a compositional approach, translating "Great Ming" (→ 明朝) and "Code" (→ 法典) separately, resulting in the descriptive but non-standard phrase "明朝法典" (Ming Dynasty Code).

In contrast, EA-RLVR's reasoning trace explicitly triggers a hypothesis check: *"Great Ming Code is likely referring to the* 大明律*... I should confirm the standard translation."*. By treating the phrase as a rigid proper noun rather than a translatable sentence fragment, EA-RLVR successfully retrieves the historiographically correct term "大明律".

Analogical Reasoning for Cultural Conventions (Case 2). Case 2 illustrates how EA-RLVR leverages parametric knowledge for style transfer. For the film title "Once Upon a Time in Venezuela", Qwen3 defaults to a dictionary translation of the idiom "Once Upon a Time" (→ 很久很久以前), missing the cinematic context. EA-RLVR, however, employs **analogical reasoning**. The thinking trace reveals a crucial intermediate step: it recalls a prototype entity, *"Once Upon a Time in America is translated as* 美往事*"* and applies this naming convention to the target entity, synthesizing the culturally attuned title "委瑞拉往事" (Venezuela Chronicles/Past). This demonstrates the model's ability to map new entities to existing cultural schemas.

Domain-Specific Disambiguation (Case 3 & 4). Polysemy poses a major challenge in entity translation. In Case 3, Qwen3 fails to resolve the term "Mahavira Hall" within a Buddhist context. Confused by the association of "Mahavira" with Jainism, it resorts to a phonetic transliteration "哈拉". EA-RLVR correctly identifies the domain constraints: *"In Chinese Buddhist terminology... Mahavira is often transliterated as* 大雄*"*. It successfully navigates the semantic shift of "Mahavira" (Great Hero) in Buddhism to produce the correct temple hall name "大雄殿". Similarly, in Case 4 (French), EA-RLVR shows a willingness to localize media titles ("Emi magique") rather than leaving them in English, reflecting a deeper adherence to target-language publication norms.

Summary of Learned Reasoning Pattern. Across these cases, a consistent meta-strategy emerges. The standard SFT/Base model tends to optimize for $P(y|x)$ at the token level, favoring high-frequency phrases (fluency) over factual precision. EA-RLVR, driven by the entity-anchored reward, learns a **"Pause-Retrieve-Verify"** loop. It (1) detects potential cultural entities, (2) suspends immediate translation to search its parametric memory for domain equivalents, and (3) verifies the candidate against the target cultural context. This shift from *translating meaning* to *matching entities* is the core driver of the observed performance gains.

Qwen3

User: Translate the follow text to Chinese (Traditional): *What was the purpose of the Great Ming Code?*

LLM: Okay, the user wants me to translate "What was the purpose of the Great Ming Code?" from English to Traditional Chinese. Let me start by understanding the original sentence. The key terms here are "Great Ming Code" and "purpose."

First, "Great Ming Code" refers to the legal code of the Ming Dynasty. In Chinese, the Ming Dynasty is called 明朝 (Míngcháo), and "Code" in this context is likely 法律 (fǎlǜ) or 法典 (fǎdiǎn). Since it's a specific historical code, using 法典 might be more appropriate. So, "Great Ming Code" would be 明朝法典 (Míngcháo Fǎdiǎn).

Putting it all together, the translation would be "明朝法典的目的是什麼？" (Míngcháo Fǎdiǎn de mùde shì shénme?). Let me double-check if there's a more standard term for "Great Ming Code." Sometimes, historical documents might have specific names, but since the user used "Great Ming Code," I think 明朝法典 is accurate.

[...]

I don't see any issues with this translation. It's clear and directly translates the original question without adding or omitting information. The user probably needs this for academic or historical purposes, so accuracy is key. No need for additional explanations, just the translated text as requested.

明朝法典的目的是什麼?

EA-RLVR

User: Translate the follow text to Chinese (Traditional): *What was the purpose of the Great Ming Code?*

LLM: Okay, the user wants me to translate "What was the purpose of the Great Ming Code?" from English to Traditional Chinese. Let me start by breaking down the sentence.

First, "What was the purpose" – in Chinese, that would be "其目的是什麼" or "目的是什麼". Then "of the Great Ming Code". Now, "Great Ming Code" is a specific term. I need to make sure I get the correct translation for that. The "Great Ming Code" is likely referring to the "大明律", which is the legal code of the Ming Dynasty. So putting it all together, the sentence would be "大明律的目的是什麼?".

Wait, let me double-check. The user might be referring to the "Great Ming Code" as "大明律" or maybe "大明律例". I should confirm the standard translation. Yes, "大明律" is the correct term for the Ming Dynasty's legal code. So the translation should be "大明律的目的是什麼?".

I should ensure that the grammar is correct in Traditional Chinese. The structure "What was the purpose of..." translates to "目的是什麼" placed after the subject. So the final translation is accurate.

大明律的目的是什麼?

Figure 8: Case study 1: Canonicalization of Historical Terminology.

Qwen3	EA-RLVR
User: Translate the follow text to Chinese (Traditional): *What is the genre of the film 'Once Upon a Time in Venezuela'?*	**User**: Translate the follow text to Chinese (Traditional): *What is the genre of the film 'Once Upon a Time in Venezuela'?*
LLM: Okay, the user wants me to translate the sentence "What is the genre of the film 'Once Upon a Time in Venezuela'?" into Chinese (Traditional). Let me start by breaking down the original sentence. First, "What is the genre of the film..." translates to "電影的類型是什麼？" in Chinese. Then, the title "Once Upon a Time in Venezuela" needs to be translated. The phrase "Once Upon a Time" is a common opening for fairy tales, so in Chinese, it's usually translated as "很久很久以前". However, since this is a film title, I should check if there's an official translation. If not, I'll use the direct translation. The country name "Venezuela" in Chinese is "委內瑞拉". So putting it all together, the title would be "很久很久以前在委內瑞拉". But wait, sometimes film titles might keep the original English title or use a slightly different translation. For example, some movies use "Once Upon a Time in..." as part of their title, so maybe it's better to keep the original English title in quotes and add the translated part. However, the user specified to translate the entire text, so I should translate the title as well. So the full translation would be: "電影《很久很久以前在委內瑞拉》的類型是什麼？" Let me double-check the grammar. The structure in Chinese is correct: "電影《...》的類型是什麼？". The title is in quotes with the Chinese translation. I think that's accurate. No need to add any extra information since the user only wants the translated text. Alright, that should be it. 電影《很久很久以前在委內瑞拉》的類型是什麼？	**LLM:** Okay, the user wants me to translate the question "What is the genre of the film 'Once Upon a Time in Venezuela'?" from English to Traditional Chinese. Let me start by breaking down the sentence. First, "What is the genre" translates to "這部電影的類型是什麼". Then, "of the film" would be "電影". So putting that together, it's "這部電影的類型是什麼". Now, the title "Once Upon a Time in Venezuela" needs careful handling. In Chinese, movie titles are often translated directly but with some cultural adaptation. "Once Upon a Time in... " is a common phrase, similar to "Once Upon a Time in..." which in Chinese is usually translated as "某年某月某日" or more literally "很久很久以前". However, for movie titles, it's better to check the standard translation. **Wait, I think "Once Upon a Time in America" is translated as "美國往事", so following that pattern, "Once Upon a Time in Venezuela" would be "委內瑞拉往事". Let me confirm. Yes, that makes sense. So combining all parts, the full translation would be "電影《委內瑞拉往事》的類型是什麼".** I should make sure the title is correctly translated and the sentence structure is natural in Chinese. Double-checking the grammar, it seems correct. The user asked only for the translated text, so I'll present that. 電影《委內瑞拉往事》的類型是什麼

Figure 9: Case study 2: Analogical Reasoning for Cultural Conventions.

Qwen3	EA-RLVR
User: Translate the follow text to Chinese (Traditional): *What is the main purpose of a Mahavira Hall in Buddhism?*	**User**: Translate the follow text to Chinese (Traditional): *What is the main purpose of a Mahavira Hall in Buddhism?*
LLM: Okay, the user is asking about the main purpose of a Mahavira Hall in Buddhism. First, I need to make sure I understand what a Mahavira Hall is. Mahavira refers to the 24th Tirthankara of Jainism, not Buddhism. Wait, the user mentioned Buddhism, but Mahavira is a key figure in Jainism. Maybe there's a confusion here. Let me verify. In Buddhism, there are different halls like the Buddha Hall, Bodhi Hall, or Dharma Hall. The term "Mahavira Hall" doesn't ring a bell in Buddhist contexts. Perhaps the user mixed up Jainism and Buddhism? Mahavira is central to Jainism, so maybe they meant a Jain temple or hall. But the question specifically says Buddhism. [···] So, the original question is in English: "What is the main purpose of a Mahavira Hall in Buddhism?" Translating that to zh_TW would be "佛教中馬哈維拉廳的主要目的是什麼？ " But I need to ensure that "Mahavira Hall" is correctly translated. Since Mahavira is a proper noun, it's usually transliterated. In Chinese, it's often written as 馬哈維拉. So the translation should be accurate. However, **I should note that in Buddhism, there isn't a Mahavira Hall**. But since the user is asking for the translation, not the correction, I'll provide the translated text as is. The answer might need to address the possible confusion, but the user only wants the translation. So the final answer is the translated question. 佛教中馬哈維拉廳的主要目的是什麼?	**LLM:** Okay, the user wants me to translate "What is the main purpose of a Mahavira Hall in Buddhism?" from English to Traditional Chinese. Let me start by breaking down the sentence. First, "What is the main purpose" – in Chinese, that would be "主要目的是什麼". Then "of a Mahavira Hall" – I need to make sure about the term "Mahavira Hall". **Mahavira is a name, right? In Buddhism, Mahavira refers to the 24th Tirthankara, so the term might be translated as "大雄寶殿" or "摩訶維拉殿"?** Wait, maybe I should check the standard translation. Wait, "Mahavira" is often transliterated as "大雄" in Chinese Buddhist contexts. So "Mahavira Hall" would be "大雄寶殿". Putting it all together, the sentence would be "佛教中大雄寶殿的主要目的是什麼？ ". Let me double-check. "Mahavira Hall" – yes, in Chinese Buddhist terminology, it's commonly known as "大雄寶殿", which is the main hall in a temple where the Buddha is enshrined. So the translation should be correct. 佛教中大雄寶殿的主要目的是什麼?

Figure 10: Case study 3: Domain-Specific Disambiguation .

Qwen3	EA-RLVR
User: Translate the follow text to French: *Who is the author of the Magical Emi, the Magic Star book series?* **LLM:** Okay, the user wants me to translate the sentence "Who is the author of the Magical Emi, the Magic Star book series?" from English to French. Let me start by breaking down the sentence. First, "Who is the author" translates to "Qui est l'auteur". Then, "of the Magical Emi" would be "de Magical Emi". Wait, should "Magical Emi" be translated or kept as is? Since it's a title, probably keep it in English. So "de Magical Emi". Next part: "the Magic Star book series". "Magic Star" is likely a title too, so keep it as "Magic Star". "Book series" is "série de livres". So putting it together: "la série de livres Magic Star". [···] Wait, the original says "the Magic Star book series". So maybe "la série de livres Magic Star" is accurate. Also, the titles are in English, so they should remain as is. The user might be referring to a specific book series, so keeping the titles in English is important for recognition. Double-checking the translation: "Qui est l'auteur de Magical Emi, la série de livres Magic Star ?" Yes, that seems correct. No grammar issues. The user just needs the translated text, so no extra explanations. Alright, that's the final translation. Qui est l'auteur de Magical Emi, la série de livres Magic Star ?	**User**: Translate the follow text to French: *Who is the author of the Magical Emi, the Magic Star book series?* **LLM:** Okay, the user wants me to translate "Who is the author of the Magical Emi, the Magic Star book series?" from English to French. Let me start by breaking down the sentence. First, "Who is the author" translates to "Qui est l'auteur". Then, "of the Magical Emi" – I need to check the title. "Magical Emi" might be a proper noun, so I should check if it's translated as "Emi magique" or maybe "L'émie magique". **Wait, the original title is "Magical Emi", so maybe "Emi magique".** Next, "the Magic Star book series". "Magic Star" as a title would be "Étoile magique" or "Étoile magique". But since it's a series, maybe "série de livres Étoile magique". Putting it all together: "Qui est l'auteur de la série de livres Emi magique, Étoile magique?" Wait, but the original says "the Magic Star book series", so maybe "la série de livres Étoile magique". But I need to check if the titles are translated correctly. Maybe "Magical Emi" is "Emi magique" and "Magic Star" is "Étoile magique". So the full translation would be "Qui est l'auteur de la série de livres Emi magique, Étoile magique?" Let me double-check the titles to make sure they're correctly translated. I think that's right. Qui est l'auteur de la série de livres Emi magique, Étoile magique?

Figure 11: Case study 4: Domain-Specific Disambiguation .