TAPO: Translation Augmented Policy Optimization for Multilingual Reasoning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable proficiency in English mathematical reasoning, yet a significant performance disparity persists in multilingual contexts, largely attributed to deficiencies in language understanding. To bridge this gap, we introduce Translation-Augmented Policy Optimization (TAPO), a novel reinforcement learning framework built upon GRPO. TAPO enforces an explicit alignment strategy where the model leverages English as a pivot and follows an understand-then-reason paradigm. Crucially, we employ a step-level relative advantage mechanism that decouples understanding from reasoning, allowing the integration of translation quality rewards without introducing optimization conflicts. Extensive experiments reveal that TAPO effectively synergizes language understanding with reasoning capabilities and is compatible with various models. It outperforms baseline methods in both multilingual mathematical reasoning and translation tasks, while generalizing well to unseen languages and out-of-domain tasks.

1 Introduction

Large language models (LLMs) (Hurst et al., 2024; Comanici et al., 2025; DeepSeek-AI et al., 2024; Yang et al., 2025) have achieved great progress in English tasks like mathematical reasoning (Jaech et al., 2024; Guo et al., 2025), even surpassing humans in competitions (Liu et al., 2025), while the multilingual disparity still exists (Huang et al., 2025b, 2023). Driven by the imbalance in the quantity and quality of training corpora across different languages, LLMs exhibit uneven language-specific capabilities including understanding and generation. With respect to the multilingual mathematical reasoning task, Kang et al. (2025) find that the multilingual gap of large reasoning models mainly stems from the language understanding.

A promising strategy (Zhu et al., 2024; She et al., 2024) to bridge this gap is multilingual alignment, where LLMs leverage a dominant language (typically English) as a pivot for processing low-resource languages. The model maps the input into the dominant language's representation for reasoning, and maps the generated output back to the original source language. As the multilingual understanding is still a bottleneck, we mainly focus on understanding and reasoning in English in this work for investigating the foundation of alignment. While some previous works (Li et al., 2024a; Bu et al., 2025) try to align the language representations directly, they have a high risk of catastrophic forgetting when applied to post-trained models, and lack the human-understandable interpretability of alignment. Other works (Yoo et al., 2025; Wang et al., 2025) train models with code-switching data in the (continual) pre-training stage, yet they still cannot fully leverage the capabilities of post-trained models.

To address the above issues, we propose a novel approach called **T**ranslation-**A**ugmented **P**olicy **O**ptimization (TAPO). We build our algorithm upon the online reinforcement learning (RL) method GRPO (Shao et al., 2024) to alleviate the catastrophic forgetting (Shenfeld et al., 2025). We adopt an explicit alignment strategy in which the model first generates an English representation of its understanding of the problem, followed by reasoning conducted in English. This explicit separation between the understanding and reasoning stages decouples the two processes, enabling more effective guidance and control during training. Intuitively, translating the input into English can be viewed as a manifestation of the model's comprehension, given that English serves as the dominant language. Consequently, we can leverage existing translation metrics to provide explicit rewards for the understanding stage, thereby sharpening the model's semantic understanding. Furthermore, to

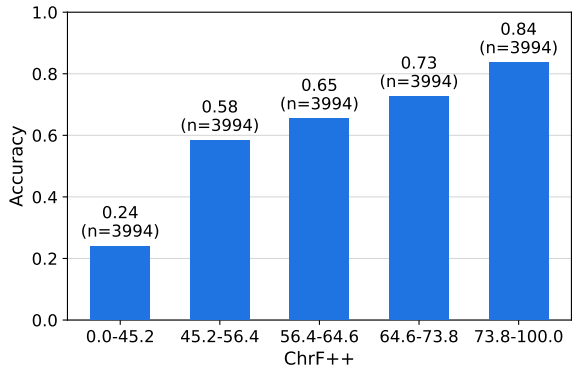

Figure 1: Accuracy grouped by ChrF++ score intervals. The scores are calculated from the outputs of Qwen2.5-3B-Instruct prompted with Translate-Test. The data is divided into five equal-sized bins (n=3994).

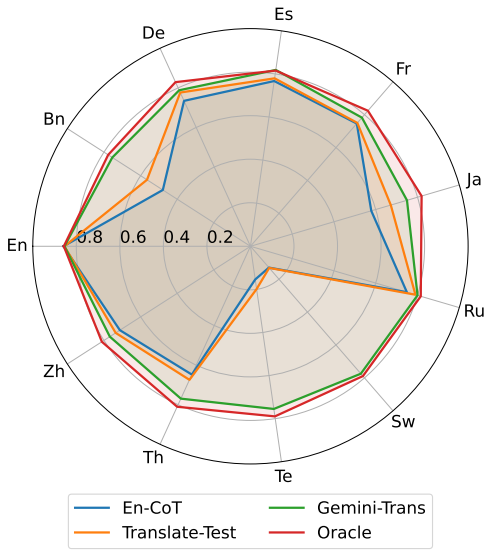

Figure 2: The performance of Qwen2.5-3B-Instruct on MGSM with different prompts. The full prompts are in Appendix A. The last two are the same as Translate-Test, but the model's translation is replaced with Gemini2.5-Flash's and the reference, respectively.

effectively integrate these signals without causing reward conflicts, we implement a step-level relative advantage strategy within the GRPO framework that calculates advantages for translation and reasoning tokens separately.

Extensive experiments are conducted on both in-domain and out-of-domain multilingual mathematical benchmarks across two different models. We also evaluate the understanding capability by assessing the translation quality of multilingual problems. Results show that our proposed method can synergize both the understanding capability and the reasoning capability, exhibiting better performance on both tasks compared to naive GRPO. Further analysis indicates that our model can generate more faithful Chain-of-Thought (CoT) process. We also provide some interesting findings in post-training the multilingual LLMs.

The main contributions of our work can be summarized as follows:

- We propose a novel method TAPO that jointly optimizes the multilingual understanding and reasoning capabilities.
- Experiments show that our method achieves best performance on both reasoning and translation tasks, and generalizes effectively to untrained languages and out-of-domain tasks.
- Our analyses provide deeper insights into the multilingual alignment and multilingual post-training strategies.

2 Preliminary Studies

2.1 Multilingual Understanding Bottleneck

To investigate the impact of language understanding on multilingual reasoning, we first examine the relationship between the model's understanding level of the problem and its reasoning accuracy. We quantify the model's understanding level of a problem by evaluating the model's translation performance based on Translate-Test prompt (shown in Table 6), where translation quality is measured using the ChrF++ score (Popović, 2017). The results in Figure 1 show that higher understanding levels correspond to higher reasoning accuracy, indicating that improving a model's language understanding capability is a key factor for enhancing multilingual reasoning.

Furthermore, we examine the extent of the multilingual understanding bottleneck in LLMs for multilingual reasoning. As shown in Figure 2, Translate-Test, in which the model self-translates the problem into English before reasoning, yields only marginal gains compared to directly providing multilingual input and requiring an English response (En-CoT). Its performance remains far below that achieved using stronger model translations (Gemini-Trans) or standard translations (Oracle), indicating that multilingual understanding remains a key bottleneck for further reasoning.

2.2 Limited Improvement with Naive GRPO

A feasible way to train multilingual reasoning is applying GRPO (Shao et al., 2024) directly with multilingual data, similar to Huang et al. (2025a). The objective function of GRPO we used in this

work is

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q,o_i\sim\pi_{\theta_{old}}}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\right.$$
$$\left.\min\left(r_{i,t}\hat{A}_{i,t}, \text{clip}(r_{i,t}, 1-\varepsilon_l, 1+\varepsilon_h)\hat{A}_{i,t}\right)\right], \quad (1)$$

where $r_{i,t}$ is the importance sampling ratio, $\hat{A}_{i,t}$ is the advantage, and q is a problem. Following DAPO (Yu et al., 2025), we remove the KL divergence loss and apply the techniques of Clip-Higher and Token-Level Policy Gradient Loss. The clipping thresholds are $\varepsilon_l = 0.2$ and $\varepsilon_h = 0.28$.

We train Qwen2.5-3B-Instruct by GRPO with the Swahili subset of MGSM8KInstruct (Chen et al., 2024). Although the reasoning performance increases steadily, achieving 38.95% accuracy (compared to 12.95% before training) in MGSM-sw, we find that the model can generate an unfaithful reasoning process while still obtaining the correct answer. As shown in the example below, the trained model almost completely misunderstands the problem. However, it predicts the numbers and the logical relationships between them correctly.

Problem: Joho hutumia komeo 2 za ufumwele wa buluu na nusu ya kiasi hicho cha ufumwele mweupe. Huwa inatumia jumla ya komeo ngapi? (*A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?*)
Output: ... The problem states that there are 2 cages for the birds and half as much space for the pet birds. We need to find the total number of cages used....\boxed{3}

We argue that training by naive GRPO on multilingual data can only improve the understanding capability to a very limited extent. Additional learning signals are necessary to enhance understanding.

3 Methodology

In this section, we will present our novel method, TAPO, to tackle the aforementioned problems. Our main idea is to enhance the comprehension of the problem by introducing a new rewards that quantifies the level of understanding, in conjunction with the existing reasoning reward. This raises two fundamental questions: How should the reward for understanding be designed? (§ 3.1) How to utilize the reward in the current RL pipeline? (§ 3.2)

3.1 Reward Modeling

Given the challenges associated with extracting the understanding component and quantifying its level, we regard the English translation of the problem as a surrogate for understanding, because most LLMs are proficient in English. And then we can use translation metrics to measure its quality. We opt to require the model to translate the question initially by using the Translate-Test prompt (Table 6).

Format Reward. The rollout trajectory o of each sample is the concatenation of the translation of the problem τ_{trans} and the reasoning process τ_{reason} containing the final answer, denoted as $o = [\tau_{trans}, \tau_{reason}]$. According to the prompt, the translation should be enclosed by `<english_translation>` and `</english_translation>` for ease of extraction. Thus, the format reward is defined as:

$$R^{format} = \begin{cases} 1, & \text{if the translation exists} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

We do not require the model to generate a CoT for the translation because there is no strong evidence that CoT can help translation (Wu et al., 2025).

Translation Reward. Following MT-R1-Zero (Feng et al., 2025b), we use the translation metrics as the translation reward. We choose ChrF++ (Popović, 2017) as the n-gram based metric and XCOMET-XL (Guerreiro et al., 2024) as the model-based metric. However, we find that xCOMET is very easy to be hacked in Swahili and it is not trained low-resource languages such as Telugu on evaluation tasks. As a result, we develop an adaptive metric according to the source language. In experiments, the adaptive metric use ChrF++ for Swahili and Telugu, and XCOMET for other languages.

$$R^{trans} = M(\tau_{trans}, q, q_{en}) \times R^{format}, \quad (3)$$

where q and q_{en} are the non-English problem and the corresponding English problem. M represents the metric function that can be ChrF++, xCOMET or the adaptive one.

Reasoning Reward. We use the outcome correctness as the reasoning reward, which has proven its effectiveness in mathematics (Guo et al., 2025).

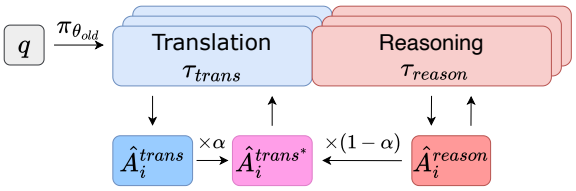

Figure 3: The illustration of the calculation process of advantages.

We evaluate the correctness by `Math-Verify`[1].

$$R^{reason} = \begin{cases} 1, & \text{if correct} \\ 0, & \text{if wrong or } R^{format} = 0 \end{cases}. \quad (4)$$

3.2 Step-level Relative Advantage

One naive approach to utilizing the rewards is to add them all together, and then directly apply GRPO. However, we identify a problem of reward conflict if we directly aggregate the rewards by addition, which hinders the reasoning. Imagine that if all trajectories in a group are right or wrong, the advantages are solely determined by the translation rewards which are real numbers. In this scenario, the advantages of half of the reasoning traces τ_{reason} possess the incorrect signs. Conversely, low-quality translations are likely to receive positive advantages if the answer is guessed correctly.

Thanks to the explicit separation of the translation step and the reasoning step, we can apply the step-level relative advantages separately, as shown in Figure 3. Inspired by Zhang et al. (2025b) and Feng et al. (2025a), the translation rewards and reasoning reward are grouped separately, and the advantages within each group are subsequently computed.

$$\hat{A}_i^{trans} = \frac{R_i^{trans} - \text{mean}(\{R_i^{trans}\}_{i=1}^G)}{\text{std}(\{R_i^{trans}\}_{i=1}^G)}$$
$$\hat{A}_i^{reason} = \frac{R_i^{reason} - \text{mean}(\{R_i^{reason}\}_{i=1}^G)}{\text{std}(\{R_i^{reason}\}_{i=1}^G)} \quad (5)$$

where $\hat{A}_i^{trans}$ is the advantage of tokens in the translation step, and $\hat{A}_i^{reason}$ is the advantage of tokens in the reasoning step. Furthermore, to circumvent the reward hacking of translation metrics and provide more signals, the final translation advantage is derived as the weighted average of the translation advantage and the reasoning advantage:

$$\hat{A}_i^{trans^*} = \alpha \hat{A}_i^{trans} + (1-\alpha)\hat{A}_i^{reason} \quad (6)$$

where $\alpha \in [0, 1]$.

[1] https://github.com/huggingface/Math-Verify

4 Experiments

4.1 Experimental Setup

Base Models. We verify our method on Qwen2.5-3B-Instruct (Qwen Team, 2024) and Llama3.2-3B-Instruct (Dubey et al., 2024). For simplicity, we refer to them as Qwen and Llama in the following paper.

Datasets. We mainly use MGSM8KIntruct (Chen et al., 2024) as the training set. Besides, we translate the training set of GSM8K (Cobbe et al., 2021) into Telugu using Gemini2.5-Flash, using the same prompt in Huang et al. (2025a). We select 5 languages to train each model: Bengali (Bn), Japanese (Ja), Swahili (Sw), Telugu (Te), Thai (Th) for Qwen, and Bn, German (De), Sw, Th, Chinese (Zh) for Llama. The number of samples in each language is about 7.4k.

Benchmarks and Metrics. To evaluate the mathematical reasoning capability, we use MGSM (Shi et al., 2023) as the in-domain testset. Note that there are eleven languages in MGSM, so 6 languages are not seen in our training data. We also include MMATH (Luo et al., 2025) and MSVAMP (Chen et al., 2024) as the out-of-domain (OOD) testsets. We evaluate each sample eight times and compute the average accuracy. We also assess the translation quality by leveraging non-English problems from MGSM as sources and English problems as references. Each source is also translated eight times and we compute the average score. The inference temperature is set to 0.6 for all tasks. Since ChrF++ and XCOMET have been used as rewards in training, LLM-as-a-judge with Gemini2.5-Flash is employed to evaluate translations (Lavie et al., 2025), denoted as Gemini Score. The prompt (Appendix A) ask the model to generate a score ranging from 0 to 100.

Baselines.

- **Base**: The original instruction models.
- **SFT-TransTest**: We use the same multilingual training data mentioned above to build the SFT dataset. The output is similar to the format of Tranlation-Test, where each answer concatenates the English translation and the English reasoning process from the original GSM8K training set.
- **QAlign** (Zhu et al., 2024): QAlign proposes a two-step SFT pipeline, including a question alignment stage and a response alignment

Models	Trained Languages					Avg	Untrained Languages						Total
	Bn	Ja	Sw	Te	Th		De	En	Es	Fr	Ru	Zh	Avg
Qwen	56.5	67.2	13.2	19.9	67.5	44.8	77.7	85.5	78.0	75.0	79.0	73.6	44.8
SFT-TransTest	1.2	0.7	1.0	1.1	0.9	1.0	0.9	7.2	1.0	0.9	1.1	1.0	1.5
QAlign	28.4	36.9	10.5	12.5	41.6	25.9	50.8	64.5	56.8	53.2	54.0	55.7	42.2
GRPO-EnCoT	75.2	**78.7**	39.0	51.0	82.2	65.2	84.0	89.4	84.7	82.7	**85.8**	**81.4**	75.8
GRPO-TransTest	**75.7**	78.2	40.9	51.2	**82.9**	65.8	84.6	**90.5**	**87.2**	**83.9**	85.4	80.6	76.4
TAPO-xCOMET	74.1	77.5	43.9	52.2	80.9	65.7	84.5	89.6	84.6	81.5	85.3	80.8	75.9
TAPO-ChrF++	74.3	76.6	48.5	57.5	80.5	67.4	**84.7**	89.9	83.4	82.0	83.8	78.6	76.3
TAPO-Adapt	73.7	75.8	**53.0**	**58.2**	80.5	**68.2**	82.2	89.1	84.1	80.2	84.3	80.1	**76.5**

Table 1: The performance of each model based on Qwen2.5-3B-Instruct.

Models	Trained Languages					Avg	Untrained Languages						Total
	Bn	De	Sw	Th	Zh		En	Es	Fr	Ja	Ru	Te	Avg
Llama	34.7	56.8	27.1	48.7	54.6	44.4	73.6	61.6	56.7	44.1	58.3	31.0	49.7
SFT-TransTest	45.3	45.6	49.1	44.5	50.9	47.1	55.4	64.8	59.4	54.2	54.8	53.3	52.5
QAlign	30.2	29.1	31.8	27.4	38.2	31.3	43.5	57.5	47.0	44.1	43.1	41.6	39.4
GRPO-EnCoT	66.2	76.7	68.8	71.3	73.0	71.2	86.7	79.7	76.4	69.2	76.6	60.4	73.2
GRPO-TransTest	67.6	79.7	70.3	74.8	75.6	73.6	86.0	81.5	77.8	70.0	78.9	64.4	75.1
TAPO-xCOMET	68.7	78.7	71.7	**76.1**	77.2	74.5	**87.6**	81.4	77.6	71.4	79.3	**66.3**	76.0
TAPO-ChrF++	**69.7**	**80.0**	**72.1**	73.8	**79.7**	**75.0**	87.3	**84.6**	**78.7**	**72.0**	78.6	64.3	**76.4**
TAPO-Adapt	68.4	79.4	71.8	75.3	77.0	74.3	87.4	81.3	78.6	70.3	**79.0**	64.5	75.7

Table 2: The performance of each model based on Llama3.2-3B-Instruct.

stage. We use the same multilingual data to build the translation training set, and the original GSM8K training set as the reasoning task.

- **GRPO-EnCoT**: We use directly GRPO train the models with En-CoT prompt. Only the reasoning reward is given.
- **GRPO-TransTest**: Same as the GRPO-EnCoT, but training with the Translation-Test prompt.

All baseline models are evaluated using the training prompts, while Base models directly use the Translation-Test prompt.

Training details. For RL training, we use verl (Sheng et al., 2024) to conduct experiments. The learning rate is set to 1e-6. The global batch size is 256 and the mini batch size is 64. The maximum response length is set to 2048 and the rollout temperature is 1.0. We train Qwen2.5-3B-Instruct and Llama3.2-3B-Instruct for 5 epochs and 2 epochs, respectively. The hyper-parameter α is set to 0.5 for training Qwen, and 0.25 for training Llama. For SFT training, we deploy LlamaFactory (Zheng et al., 2024). The learning rate is set to 1e-6, with 3% linear warm-up steps, followed by cosine decay to 1e-7. The global batch size is 64.

4.2 Main results

TAPO demonstrates superior performance in trained languages and comparable performance in untrained languages. Table 1 and Table 2 demonstrate the results on MGSM. In trained languages, the Qwen trained by TAPO-Adapt achieve the best average scores. The main contributions come from the two low-resource languages, Swahili and Telugu, increasing by 12.1 and 7.0 compared to GRPO-TransTest, respectively. Its performance in untrained languages is a bit lower than GRPO, but the gaps are minor. The Llama model trained by TAPO-ChrF++ demonstrate more substantial performance gain in both trained and untrained languages. The effectiveness in trained languages suggests that enhancing the understanding capability is indeed beneficial for downstream reasoning, especially in low-resource languages.

TAPO enhances the translation performance more effectively, indicating better understanding capabilities. We evaluate the problem translation mainly using the Gemini Score. The detailed Chrf++ scores can be found in Appendix B. The translations are extracted from the whole trajectory if using the Translate-Test prompt, or obtained through a separate translation prompt. The results in Table 3 demonstrate that the proposed method enhances the translation quality substantially across nearly all languages, including the untrained languages. Qwen with TAPO-Adapt achieves the highest average score of 75.86, a significant improvement over the GRPO baselines. Particularly,

Models	Bn	De	Es	Fr	Ja	Ru	Sw	Te	Th	Zh	Avg
Qwen	58.60	84.07	87.30	85.89	81.10	87.28	7.72	17.82	77.68	86.72	67.42
GRPO-EnCoT	62.97	83.58	87.88	86.66	80.26	88.88	18.18	29.31	77.99	83.66	69.94
GRPO-TransTest	62.59	82.69	86.76	85.89	79.91	86.85	19.94	33.67	79.09	83.87	70.12
TAPO-xCOMET	68.70	79.73	82.20	81.00	83.11	**89.64**	2.74	43.93	80.72	**87.58**	69.93
TAPO-ChrF++	68.79	86.30	**90.18**	87.55	82.54	89.03	**34.30**	**47.89**	81.76	85.88	75.42
TAPO-Adapt	**70.67**	**87.98**	90.16	**88.38**	**84.12**	89.03	33.17	45.49	**83.84**	85.79	**75.86**
Llama	59.59	74.88	81.14	76.92	59.90	76.75	50.17	55.55	59.44	65.67	66.00
GRPO-EnCoT	56.90	64.54	71.26	66.40	53.91	64.92	47.86	54.32	54.96	57.92	59.30
GRPO-TransTest	66.38	78.76	81.44	81.30	69.55	78.35	54.99	62.74	70.76	75.58	71.98
TAPO-xCOMET	72.52	84.60	87.86	85.30	74.71	84.28	61.84	68.49	76.13	81.87	77.76
TAPO-ChrF++	**72.89**	**85.87**	89.25	85.60	**75.96**	**85.77**	63.62	68.28	77.30	**83.71**	**78.82**
TAPO-Adapt	72.82	85.18	**89.32**	**85.84**	75.41	85.31	**63.72**	**69.61**	**78.28**	82.62	78.81

Table 3: The Gemini Scores of each model on the task of translating the MGSM problems from non-English to English. The score of each sample is averaged by eight random runs.

Models	MMATH	MSVAMP
Qwen	54.25	77.55
GRPO-EnCoT	55.36	77.28
GRPO-TransTest	**57.63**	76.39
TAPO-Adapt	55.82	**78.41**
Llama	28.07	61.43
GRPO-EnCoT	42.22	73.75
GRPO-TransTest	42.34	76.02
TAPO-ChrF++	**43.78**	**77.25**

Table 4: The performance of each model on OOD tasks. The scores are averaged across all the languages in the benchmarks.

the improvements of Qwen in lower-resource languages such Sw and Te, are more than 14 points compared to baselines. One exception is that the performance of Qwen trained by TAPO-xCOMET in Sw deteriorates due to the fragility of the metric. Interestingly, although not receiving any translation rewards, GRPO-TransTest increases the translation quality slightly. This indicates that the reasoning reward is also advantageous to the translation to a certain degree, which supports the design of the final translation advantage.

TAPO possesses the generalizability to OOD tasks. We choose the best performing TAPO variants and evaluate them on OOD tasks. As shown in Table 4, TAPO brings consistent improvement on most of tasks, while the Qwen's GRPO variants even exhibit performance degradation in MVSAMP. The relatively small gains from TAPO-Adapt on MMATH is likely because the benchmark contains many languages not seen during training, which aligns with our earlier findings.

SFT induces catastrophic forgetting in post-trained models. The results of SFT-TransTest and QAlign in Table 1 and Table 2 decrease the original performance of the post-trained models in many languages, especially high-resource ones. Surprisingly, the performance of the Qwen almost drops to zero after SFT-TransTest. We cross-validate the unexpected results, and also find that the translation SFT in QAlign leads to similar translation performance drop. In contrast, online methods such as GRPO and TAPO demonstrate efficacy in averting catastrophic forgetting and ensuring consistent enhancement.

5 Analysis

In this section, we provide some ablation studies along with further analysis.

5.1 Faithfulness of CoT

Despite having evaluated the translation quality, we still want to confirm the faithfulness of the CoT. We define faithfulness as the semantic alignment between the math problem and the CoT, independent of the reasoning correctness. We prompt Gemini2.5-Flash to provide a faithfulness score ranging from 0 to 5 (Appendix A). One sample of each problem is chosen to compute the score. As shown in Figure 4, TAPO can effectively improve the faithfulness of the CoT, significantly[2] better than pure GRPO variants.

5.2 Reward Conflict in the Trajectory-level Reward

We compare our step-level reward with the trajectory-level reward, which directly apply the sum of R^{trans} and R^{reason} to the whole trajectory. The results in Table 5 demonstrate the con-

[2]We use T-test with $p = 0.05$.

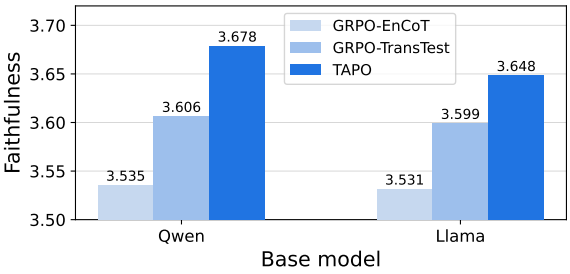

Figure 4: The average faithfulness scores across languages on MGSM. The results of TAPO are from TAPO-Adapt for Qwen and TAPO-ChrF++ for Llama.

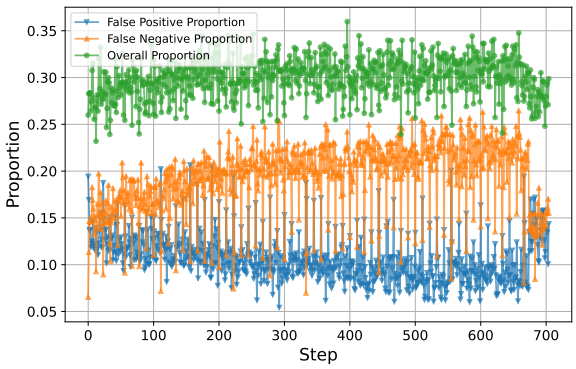

Figure 5: The proportions of false positive and false negative advantages across each training step. The overall proportion is the sum of the two proportions.

sistent superiority of the step-level reward over the trajectory-level one across different models and benchmarks. Furthermore, we identify the reward conflict problem introduced by the trajectory-level reward that hinders the effectiveness of training. The advantages normalized from rewards can have the wrong signs if the translation reward is dominant. The advantage assigned to a trajectory is considered a false positive if it is positive despite the answer being incorrect, and a false negative if it is negative despite the answer being correct. As illustrated in Figure 5, we count the proportions of false positive and false negative advantages across the training steps of Qwen with the adaptive metric. The overall conflict proportions are approximately 0.3, which may exert a negative influence on the learning process.

Type	MGSM	MMATH	MSVAMP
Qwen			
Traj-level	75.33	54.82	**79.59**
Step-level	**76.45**	**55.82**	78.41
Llama			
Traj-level	74.11	42.13	75.87
Step-level	**76.41**	**43.78**	**77.25**

Table 5: The comparison between the trajectory-level reward by addition and our step-level reward. Scores are averaged across all languages.

5.3 Impacts of Translation Metrics

The selection of translation metrics demonstrates varying degrees of influence across different languages. As shown in Table 1, TAPO-ChrF++ more often benefits the low-resource languages such as Sw and Te, while TAPO-xCOMET tends to perform better in high-resource languages like En and Ru. The model-based metric XCOMET is easier to be hacked during RL. We find that just copying the source problem in Sw as the translation can receive high XCOMET scores, yielding incorrect signals. The N-gram based metric ChrF++ is more robust in all kinds of languages, although it may fail to differentiate high-quality translations. The adaptive metric can combine the advantages of both to some extent, exhibiting comparable performance in Sw and Te with ChrF++, and in other languages with XCOMET.

5.4 Influence of the Hyper-parameter α

The Hyper-parameter α controls the weights of advantages from the translation metric and the outcome correctness. As illustrated Figure 6, the impacts of α on both tasks reveals a trade-off between reasoning capabilities and translation quality. The reasoning accuracy follows a non-monotonic trend, initially rising to achieve a global peak of 76.41% at $\alpha = 0.25$ before declining sharply to a minimum of roughly 73.22% at $\alpha = 0.75$. The overall trend is an initial increase followed by a decrease. Conversely, the Gemini Score exhibits a strong positive correlation with α, exhibiting a steep increase when α is small and a gradual rise when α exceeds 0.25. This indicates that a small amount of translation signal can yield substantial reasoning improvement, while a further escalation in translation signal will influence the learning of reasoning ability.

5.5 Convergence of Different Languages

By observing the validation results on MGSM, we find that different languages exhibit varying speed of convergence. As illustrated in Figure 7, the scores of the three mid-resource languages are converged at the early stage of training. The score of Bn achieve the plateau at the step 300, and even decreases slightly in the following training steps. On the contrary, the scores of the two low-resource lan-

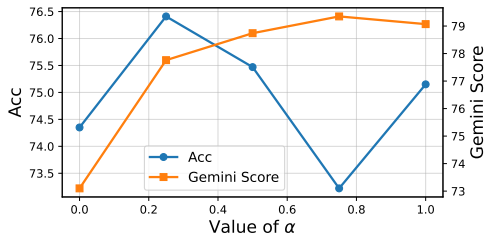

Figure 6: The impacts of α on the reasoning and the translation. Acc is the average accuracy of Llama trained by TAPO-ChrF++ on MGSM. Gemini Score evaluates the model's translation quality.

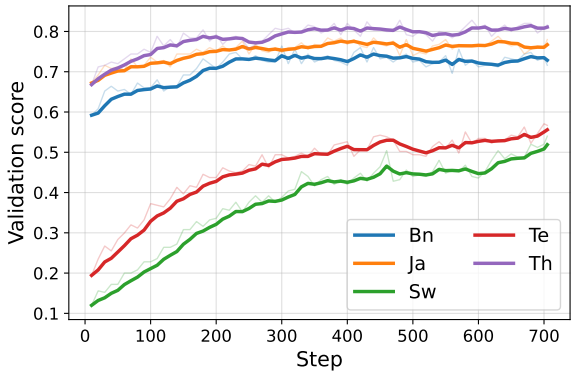

Figure 7: The validation scores of Qwen trained by TAPO-Adapt. The bold lines are smoothed by Exponential Moving Average.

guages, Sw and Te, continuously rise until the end of training, showing no convergence. This suggests that overtraining the converged languages is pointless, and more computational resources should be allocated to training low-resource languages. This can be implemented by better training data distribution or dynamic sampling (Yu et al., 2025), which we leave as future work.

6 Related Works

Multilingual Reasoning. Despite the dominant performance in English, prior studies demonstrate that LLMs retain substantial potential for multilingual reasoning (Ye et al., 2023; Kew et al., 2024). To exploit this potential, existing methods primarily enhance cross-lingual alignment through additional SFT (Zhu et al., 2024; Zhang et al., 2024) or RL (Zhang et al., 2025a; Hwang et al., 2025), aiming to improve multilingual reasoning while preserving established reasoning behaviors. However, recent analyses show that LLMs exhibit significantly stronger semantic understanding in English than in non-English languages (Park et al., 2025; Tam et al., 2025), suggesting that reasoning directly in non-English impairs performance.

In contrast to prior approaches that rely on reasoning natively, we adopt an understand-then-reason paradigm, expressing non-English inputs in English to reduce comprehension difficulty before reasoning. Similarly, Huang et al. (2023) encourage translation before reasoning via prompting, but limited by the model's intrinsic translation ability. In contrast, we explicitly train translation and reasoning with reward design.

Multilingual Alignment. To reduce the multilingual disparity, a variety of alignment methods have been proposed in recent years. Some approaches inject multilingual knowledge by training on code-switched data (Yoo et al., 2025; Wang et al., 2025), while others employ contrastive learning to map representations into a shared semantic space (Li et al., 2024b; Bu et al., 2025; Li et al., 2024a). Additional works activate language-related parameters through latent disentanglement (Zhao et al., 2025; Ye et al., 2025), or incorporate external models to assist multilingual understanding (Yoon et al., 2024; Huang et al., 2024). These methods effectively achieve multilingual representation alignment, while offering limited attention to the enhancement of reasoning behaviors.

RL has been widely validated as an effective approach for enhancing reasoning capabilities in LLMs (Shao et al., 2024; Guo et al., 2025), and its benefits extend to multilingual settings as well (Huang et al., 2025a; Lee et al., 2025; She et al., 2024). However, existing RL-based approaches largely overlook the distinction between multilingual understanding and reasoning processes. To address the limitation, we propose an RL framework that jointly optimizes multilingual understanding and reasoning, further unlocking the potential of LLMs for multilingual reasoning.

7 Conclusion

In this paper, we propose a novel method TAPO to enhance the reasoning capability. We leverage the form of explicit alignment by decoupling the understanding and reasoning process, and employ a step-level relative advantage to optimize both capabilities more effectively. Experiments demonstrate the superiority of our method on both reasoning and translation tasks with different models, generalizable to OOD tasks and untrained languages. Further analyses provide deeper insights into multilingual post-training and alignment.

Limitations

The limitations of our work are discussed below:

- We do not utilize the more advanced translation metrics like LLM-as-a-judge by strong models due to the cost and efficiency. As the metric has great influence in our framework, better metrics may bring more improvement.
- The reasoning trace and the answer are still in English, which may not be friendly to users not familiar with English.
- We do not train bigger models which have already had multilingual capabilities to a great extent. Also, we do not train the large reasoning models which may exhibit different behaviour.

References

Mengyu Bu, Shaolei Zhang, Zhongjun He, Hua Wu, and Yang Feng. 2025. AlignX: Advancing multilingual large language models with multilingual representation alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6471–6500, Suzhou, China. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025a. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*.

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

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

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

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Shulin Huang, Yiran Ding, Junshu Pan, and Yue Zhang. 2025a. Beyond english-centric training: How reinforcement learning improves cross-lingual reasoning in llms. *arXiv preprint arXiv:2509.23657*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025b. BenchMAX: A comprehensive multilingual evaluation suite for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16751–16774, Suzhou, China. Association for Computational Linguistics.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficiently boosting LLM reasoning in non-english languages. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete,

and Paul Pu Liang. 2025. Learn globally, speak locally: Bridging the gaps in multilingual reasoning. *CoRR*, abs/2507.05418.

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

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Geunbae Lee. 2025. Why do multilingual reasoning gaps emerge in reasoning language models? *arXiv preprint arXiv:2510.27269*.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning english-centric llms into polyglots: How much multilinguality is needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13097–13124. Association for Computational Linguistics.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Jungyup Lee, Jemin Kim, Sang Park, and SeungJae Lee. 2025. Making qwen3 think in korean with reinforcement learning. *CoRR*, abs/2508.10355.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024a. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.

Jiahuan Li, Shujian Huang, Aarron Ching, Xinyu Dai, and Jiajun Chen. 2024b. Prealign: Boosting cross-lingual transfer by early establishment of multilingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10246–10257. Association for Computational Linguistics.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Wenyang Luo, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2025. Mmath: A multilingual benchmark for mathematical reasoning. *arXiv preprint arXiv:2505.19126*.

Cheonbok Park, Jeonghoon Kim, Joosung Lee, Sanghwan Bae, Jaegul Choo, and Kang Min Yoo. 2025. Cross-lingual collapse: How language-centric foundation models shape reasoning in large language models. *CoRR*, abs/2506.05850.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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*.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2025. Rl's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*.

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*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. 2025. Language matters: How do multilingual input and reasoning paths affect large reasoning models? *CoRR*, abs/2505.17407.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Investigating and scaling up code-switching for multilingual language model pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11032–11046, Vienna, Austria. Association for Computational Linguistics.

Di Wu, Seth Aycock, and Christof Monz. 2025. Please translate again: Two simple experiments on whether human-like reasoning helps translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20435–20451, Suzhou, China. Association for Computational Linguistics.

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*.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *CoRR*, abs/2306.06688.

Yangfan Ye, Xiaocheng Feng, Zekun Yuan, Xiachong Feng, Libo Qin, Lei Huang, Weitao Ma, Yichong Huang, Zhirui Zhang, Yunfei Lu, Xiaohui Yan, Duyu Tang, Dandan Tu, and Bing Qin. 2025. Cc-tuning: A cross-lingual connection mechanism for improving joint multilingual supervised fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19036–19051. Association for Computational Linguistics.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2025. Code-switching curriculum learning for multilingual transfer in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7816–7836, Vienna, Austria. Association for Computational Linguistics.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7502–7522. Association for Computational Linguistics.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8037–8051. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025a. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning. *CoRR*, abs/2510.07300.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. 2025b. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. *arXiv preprint arXiv:2507.22844*.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, and 1 others. 2025. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *arXiv preprint arXiv:2505.15257*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*.

A Prompts

We present the prompts used in this section. Table 6 and Table 7 show the Translate-Test prompt and the En-CoT prompt. Table 8 is the prompt used to evaluate the translation quality. Table 9 is the prompt used to evaluate the faithfulness.

B The Detailed ChrF++ Scores

Table 10 shows the detailed ChrF++ scores.

```
[System]
Please solve the following problem by providing a
detailed, step-by-step response.

Your output **must** be structured into the following
three sections:

### 1. Translation

* Translate the provided problem into English.
* If the problem is already in English, simply
reproduce the original problem statement.
* Enclose the final English text within
`<english_translation>` and `</english_translation>`
tags.

### 2. Thought

* Present your comprehensive, step-by-step reasoning
process for arriving at the solution.
* This reasoning must be transparent, logical, and
easy to follow. Break down your process to cover
problem analysis, strategy & Planning, step-by-step
execution, and verification.

### 3. Solution

* After reasoning, provide the final, clear, and
accurate answer.
* If the result is a closed-form answer (such as a
numerical value, formula, or multiple-choice
selection), please enclose it using the `\boxed{}`
format.

[User]
{question}
```

Table 6: The prompt of Translate-Test.

```
[System]
Please solve the following problem by providing a
detailed, step-by-step response.

Your output **must** be structured into the following
two sections:

### 1. Thought

* Present your comprehensive, step-by-step reasoning
process in English for arriving at the solution.
* This reasoning must be transparent, logical, and
easy to follow. Break down your process to cover
problem analysis, strategy & Planning, step-by-step
execution, and verification.

### 2. Solution

* After reasoning, provide the final, clear, and
accurate answer.
* If the result is a closed-form answer (such as a
numerical value, formula, or multiple-choice
selection), please enclose it using the `\boxed{}`
format.

[User]
{question}
```

Table 7: The prompt of En-CoT

```
Score the following translation from {source_lang} to
{target_lang} on a scale from 0 to 100, where a score
of 0 means a broken or poor translation; 33 indicates
a flawed translation with significant issues; 66
indicates a good translation with only minor issues in
grammar, fluency, or consistency; and 100 represents
a perfect translation in both meaning and grammar.

Answer with only a whole number representing the
score, and nothing else.

{source_lang} source text: {source_seg}
{target_lang} translation: {target_seg}
{target_lang} reference: {reference_seg}
```

Table 8: The prompt of LLM-as-a-judge for translation evaluation.

```
You are a strict Cross-Lingual Logic Auditor. Your
task is to verify **Semantic Alignment** between a
math problem written in {tgt_lang} and a reasoning
trace written in English. Determine if the Reasoning
Trace is based on a **faithful understanding** of the
Problem. Do NOT evaluate if the math is correct.
Evaluate ONLY if the *translation of premises* is
accurate.

**Input:**
<Problem>
{problem}
</Problem>

<Reasoning>
{reasoning}
</Reasoning>

**Audit Checklist:**
1.  **Numerical & Entity Fidelity:** Do all
quantities, variable definitions, and actor names in
the Reasoning map exactly to the {tgt_lang} text?
2.  **Constraint Fidelity:** Are logical
relationships (e.g., "more than," "distributed
equally," "integers only") preserved?
3.  **No Information Leakage:** Did the reasoning
invent premises not found in the source text?

**Scoring Rubric:**
* **5 (Perfect):** The reasoning is based on an exact,
flawless interpretation of the {tgt_lang} premises.
* **4 (High):** Interpretation is correct, but minor
nuance in phrasing is lost (without affecting logic).
* **3 (Mixed):** Key premises are correct, but a
secondary constraint is missed or slightly altered.
* **2 (Low):** A primary constraint or numerical
value is misinterpreted.
* **1 (Critical):** Fundamental misunderstanding of
the problem's goal or setup.
* **0 (Hallucination):** The reasoning is unrelated
to the problem.

**Output:**
Return **ONLY** the integer score (0-5). Do not
include formatting, explanation, or text.
```

Table 9: The prompt for evaluating the faithfulness.

Models	Bn	De	Es	Fr	Ja	Ru	Sw	Te	Th	Zh	Avg
Qwen	56.37	70.69	73.35	65.95	60.47	68.29	34.36	41.90	59.11	63.07	59.36
GRPO-Encot	57.54	71.09	73.39	65.63	60.09	68.72	41.86	45.21	59.76	61.13	60.44
GRPO-Transtest	55.39	69.29	71.61	64.05	57.80	66.88	41.32	45.42	58.01	60.20	59.00
TAPO-xCOMET	59.04	66.37	67.72	60.71	61.40	67.97	18.38	51.48	60.73	61.87	57.57
TAPO-ChrF++	65.64	74.62	76.41	68.26	65.64	72.14	54.67	59.97	64.89	66.02	66.83
TAPO-Adapt	62.98	73.70	74.96	67.15	63.23	70.25	54.51	59.40	63.06	63.26	65.25
Llama	54.75	67.33	70.55	61.48	52.01	63.76	55.25	54.04	51.00	52.14	58.23
GRPO-Encot	50.27	56.85	60.70	51.86	45.26	51.65	48.66	49.86	45.18	41.66	50.19
GRPO-Transtest	56.08	68.25	70.35	62.30	54.76	63.20	55.95	56.12	54.13	54.81	59.59
TAPO-xCOMET	61.12	71.79	73.01	65.21	58.85	67.11	59.80	60.82	59.25	60.53	63.75
TAPO-ChrF++	65.36	75.74	76.61	67.23	62.47	70.82	65.30	65.27	63.35	64.89	67.70
TAPO-Adapt	62.45	73.44	75.35	66.05	59.93	69.12	64.59	62.57	61.39	61.73	65.66

Table 10: The Chrf++ scores of each model on the task of translating the MGSM problems from non-English to English. The score of each sample is averaged by eight random runs.