Cross-lingual Self-Consistency for Multilingual Reasoning with Language Models

Ahmed Elhady[1] **Eneko Agirre**[1] **Mikel Artetxe**[1,2]
[1]HiTZ Center, University of the Basque Country (UPV/EHU) [2]Reka AI
{ahmed.salemmohamed,e.agirre,mikel.artetxe}@ehu.eus

Abstract

Despite expanding their multilingual coverage, the advanced reasoning capabilities of LLMs remain largely confined to a few high-resource languages like English. To address this, we propose an *unsupervised* Reinforcement Learning (RL) approach to enhance multilingual reasoning by enforcing cross-lingual self-consistency—the principle that a model should produce the same final answer for equivalent problems in different languages. Existing methods are limited by the scarcity of multilingual reasoning data, and show weak generalization to unseen languages. Our approach requires neither gold answers nor parallel data, and it achieves average gains of up to **21.7%** on MGSM across 10 languages. In addition, our method demonstrates strong generalization, with an **18.2%** mean improvement on MGSM languages unseen during training, and up to **6.2%** gain on 3 out-of-distribution benchmarks. These results show the potential of consistency-based methods to improve the multilingual capabilities of LLMs without requiring supervised data. [1]

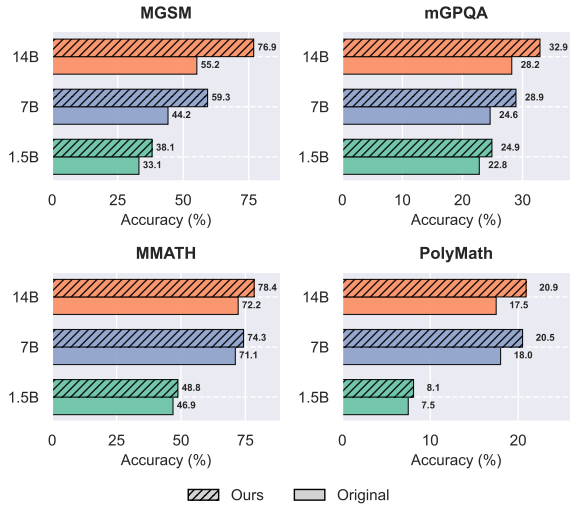

Figure 1: **Accuracy of Qwen2.5 Instruct.** Our method significantly improves over the original model, while only requiring English questions for training (without gold answers or pre-existing translations).

1 Introduction

Recent advances in Large Language Models (LLMs) show a promising trend to support a wider range of languages. For example, the pretraining of Qwen3 models (Yang et al., 2025a) included over 100 languages compared to 29 for Qwen2.5 (Yang et al., 2024). Nevertheless, their performance is limited to basic capabilities, while their effectiveness in tasks that require complex reasoning remains concentrated in a limited set of languages, such as English and Chinese (Ghosh et al., 2025).

Consequently, prior work predominantly leverages the reasoning on a strong language, typically English. For instance, given a question in the source language, the model can be prompted to reason in an English chain-of-thought instead of the native language (Liu et al., 2025; Zhu et al., 2024). As an alternative, the source-language query is translated into English in a first step, to then perform the reasoning in English in a second step. Translations can be provided by the model itself (self-translation), or, alternatively, by an external machine translation engine (Liu et al., 2025). Though simple, these strategies limit applicability in realistic scenarios where reasoning steps must be expressed in the source language.

In order to reduce the performance gap when reasoning in the source language, prior work focused on improving the cross-lingual alignment. Some works use supervised fine-tuning (SFT) on translations of questions to and from the reference languages (Zhu et al., 2024) or control the response generation by manipulating the logits to mimic those of English (Wang et al., 2025a; Huang et al., 2024). Other works use post-training methods,

[1]Models are available through this Huggingface Collection

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 43454–43466
October 24-29, 2026 ©2026 Association for Computational Linguistics

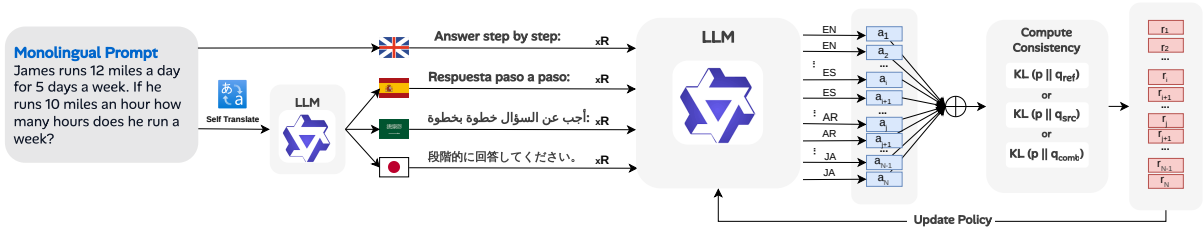

Figure 2: **Overview of our cross-lingual self-consistency framework**. Each monolingual prompt is self-translated by the model into the target languages and answered in every language, yielding set of outputs Y. Rewards are computed as the consistency between the output distribution $p(y)$ and a target distribution $q(y)$ — estimated from the reference language (q_{ref}), the prompt language (q_{src}), or all outputs combined (q_{comb}).

namely DPO (Rafailov et al., 2023), to push models to generate correct reasoning traces translated from English and discourage incorrect, natively generated ones (Yang et al., 2025b; Lee et al., 2025). These approaches rely on translated questions and reasoning traces for supervision and, as a result, generalization to unseen languages and domains is constrained by both machine-translation quality and the scarcity of high-quality manually translated data (Ghosh et al., 2025).

In this paper, we propose to use cross-lingual self-consistency as an **unsupervised** signal for RL. Our method extends existing monolingual self-consistency approaches, which capture the intuition that different reasoning paths should lead to consistent final answers, to further push the model to produce the same final answer for equivalent problems in different languages (Figure 2). This design guarantees broad practical applicability, since it requires **no access to ground truth labels nor pre-translated reasoning traces**. This is specially useful given that multilingual reasoning data is only available for evaluation, and training resources are largely monolingual (Ghosh et al., 2025). Our experimental setup only uses English questions.

We demonstrate the effectiveness of our method across three model scales, achieving an **average performance improvement of 7.8% to 21.7%** over 10 languages on the MGSM benchmark (Figure 1). The improvements can be amplified when high-quality translations for the questions are available (§6.4). Furthermore, our approach exhibits strong generalization, yielding up to **18.2% average improvement on unseen languages** within MGSM and extending effectively to other domains with a 4.7% average gain on the mGPQA (Huang et al., 2025) and 6.8% on MMATH (Luo et al., 2025) benchmarks (§6.2). Collectively, these results show that our method is effective at narrowing the multilingual reasoning gap in LLMs, advancing their robustness and equity across languages.

2 Background

Recent work has shown improvements in the monolingual reasoning capabilities of LLMs by maximizing the consistency of the model (He et al., 2026; Prabhudesai et al., 2025; Chen et al., 2025; Agarwal et al., 2025; Cui et al., 2025). We next formalize this framework, and present our proposed extension to the multilingual scenario in §3.

Given an input x, let $\{(c_i, y_i) \sim \pi(\cdot \mid x)\}_{i=1}^{N}$ denote a set of N trajectories sampled from the policy model π, where c_i is the chain-of-thought reasoning and y_i is the final answer following it. The core idea behind these methods is that, even if the reasoning paths c_i can differ, a strong model should always arrive at the same (ideally correct) final answer y_i. This principle of final answer consistency can be optimized through RL, even if the ground truth answer is not known. To that end, existing methods minimize the entropy of the marginalized answer distribution.

More precisely, let $\hat{p}_\pi(y|x)$ be the empirical probability of a specific final answer y:

$$\hat{p}_\pi(y|x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(y_i = y) \tag{1}$$

The aggregated reward over all sampled generations of x is the negative of the entropy of this distribution:

$$\begin{aligned} R_{\text{total}}(x) &= -H(\hat{p}_\pi(\cdot|x)) \\ &= \sum_{y} \hat{p}_\pi(y|x) \log \hat{p}_\pi(y|x) \end{aligned} \tag{2}$$

and the reward of each individual generation is decomposed as follows:

$$R(x, c_i, y_i) = \log \hat{p}_\pi(y_i|x) \tag{3}$$

so it holds that $R_{\text{total}}(x) = \frac{1}{N}\sum_{i=1}^{N} R(x, c_i, y_i)$. This reward function is optimized through RL, with GRPO (Shao et al., 2024) being a common choice.

3 Proposed Method

In the multilingual scenario, not only do we want models to produce consistent final answers within each language, but we also want them to produce consistent final answers for equivalent inputs in different languages. We next propose a multilingual extension of the framework in §2 that formalizes this principle of cross-lingual self-consistency.

Let $\mathbf{x} = \{x^{(l)}\}_{l=1}^{L}$ be a set of parallel inputs in L languages, and let $\{(c_i^{(l)}, y_i^{(l)}) \sim \pi(\cdot \mid x^{(l)})\}_{i=1}^{N}$ be a set of N generations for the input in language l. We assume the final answers $y_i^{(l)}$ are normalized into a shared canonical form (e.g., numeric values), so that equality is well-defined across languages. We define the following aggregated reward for the input in each language $x^{(l)}$, which replaces the entropy term in Equation 2 with the cross-entropy with respect to a second distribution $\hat{q}_\pi$:

$$\begin{aligned} R_{\text{total}}(x^{(l)}) &= -H\left(\hat{p}_\pi(\cdot|\mathbf{x}, l), \hat{q}_\pi(\cdot|\mathbf{x}, l)\right) \\ &= \sum_y \hat{p}_\pi(y|\mathbf{x}, l) \log \hat{q}_\pi(y|\mathbf{x}, l) \end{aligned} \tag{4}$$

where $\hat{p}_\pi(y|\mathbf{x}, l)$ is the empirical marginalized answer distribution in language l:

$$\hat{p}_\pi(y|\mathbf{x}, l) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(y_i^{(l)} = y\right) \tag{5}$$

and $\hat{q}_\pi(y|\mathbf{x}, l)$ is the empirical target distribution, for which we define several alternatives below. Replacing entropy with cross-entropy allows each language-specific answer distribution to be aligned to a target distribution derived from other languages, thereby operationalizing cross-lingual consistency. The reward for each sampled generation is given by

$$R\left(x^{(l)}, c_i^{(l)}, y_i^{(l)}\right) = \log \hat{q}_\pi\left(y_i^{(l)}|\mathbf{x}, l\right) \tag{6}$$

so, analogous to the monolingual case, it holds that $R_{\text{total}}(x^{(l)}) = \frac{1}{N}\sum_{i=1}^{N} R(x^{(l)}, c_i^{(l)}, y_i^{(l)})$.

In our experiments, we explore the following alternatives for $\hat{q}_\pi$, which differ on the language(s) they use to estimate the target distribution:

1. **The source language l:** In this case, the source and target distributions are the same:

$$\hat{q}_\pi(y|\mathbf{x}, l) = \hat{p}_\pi(y|\mathbf{x}, l) \tag{7}$$

Given that $H(p, p) = H(p)$, this is equivalent to minimizing the entropy of the marginalized answer distribution for each language without any constraint on cross-lingual consistency. When $L = 1$, this further becomes equivalent to the formulation in §2, making our framework a strict generalization of monolingual self-consistency.

2. **The reference language** en**:** We take the marginalized answer distribution in another (reference) language, which, without loss of generality, we set to English:[2]

$$\hat{q}_\pi(y|\mathbf{x}, l) = \max\left(\hat{p}_\pi(y|\mathbf{x}, \text{en}), \epsilon\right) \tag{8}$$

ϵ prevents the reward becoming $-\infty$ when $\hat{p}_\pi(y|\mathbf{x}, \text{en}) = 0$ and $\hat{p}_\pi(y|\mathbf{x}, l) > 0$, while still strongly favoring answers that are probable in English.[3] This variant can be seen as a form of cross-lingual distillation that, by aligning all languages to the same reference, indirectly makes them more consistent with one another.

3. **All languages combined:** We take the empirical answer distribution marginalizing over all languages and trajectories:

$$\hat{q}_\pi(y|\mathbf{x}, l) = \frac{1}{L}\sum_{k=1}^{L} \hat{p}_\pi(y|\mathbf{x}, k) \tag{9}$$

This variant directly optimizes cross-lingual self-consistency by pushing the final answer distribution in each language to match the aggregated one across all languages. Unless otherwise indicated, we use this variant in our experiments.

Our approach is **unsupervised** in that it does not require any parallel data labeled by humans. To obtain these, we adopt **self-translation**: starting from a set of English questions, we prompt the model itself to translate them into other languages. Although self-translation is noisy, we adopt it for its practicality and scalability, while avoiding external translation systems that could inject additional knowledge into the model. In practice, we find that

[2] In practice, one would choose the language that is expected to have the strongest monolingual performance which, in most scenarios (including ours), is English.

[3] We set $\epsilon = 0.01$ in our experiments, but find that it is rarely triggered in practice.

	bn	de	en	es	fr	ja	ru	sw	te	th	zh	avg
Qwen-2.5 1.5B IT	12.0	41.2	52.4	48.8	48.1	31.6	46.5	2.8	4.4	34.6	41.2	33.1
trans-test	14.4	44.2	52.4	50.0	42.1	23.6	38.8	8.4	**12.4**	36.8	40.8	33.4 (+0.3)
LIDR (Yang et al., 2025b)	18.0	52.0	**57.6**	54.4	54.0	**39.6**	**55.6**	**8.4**	6.4	37.2	**55.6**	39.9 (+6.8)
S1 (Yong et al., 2025)	17.2	47.6	54.1	57.2	52.8	32.8	50.0	6.0	4.4	40.7	50.4	37.6 (+4.5)
self-cons (ours)	**21.2**	**53.6**	50.3	**63.2**	**56.4**	36.0	**55.2**	5.2	7.2	**48.0**	54.0	**40.9** (+7.8)
Qwen-2.5 7B IT	30.0	60.4	78.6	40.2	48.2	49.8	37.4	19.8	18.2	54.8	49.2	44.2
trans-test	33.4	67.6	78.6	50.3	54.4	43.6	39.2	11.4	18.6	60.2	52.6	46.4 (+2.2)
LIDR (Yang et al., 2025b)	51.2	68.8	79.6	60.4	66.4	60.4	55.2	**25.2**	17.2	56.0	60.0	54.6 (+10.3)
S1 (Yong et al., 2025)	50.0	63.6	**88.2**	66.4	70.0	**63.2**	72.4	8.8	15.6	64.4	**60.4**	55.9 (+11.7)
self-cons (ours)	**54.8**	**75.2**	81.8	**72.4**	**71.6**	55.6	**76.0**	17.6	**24.0**	**73.2**	59.6	**59.3** (+15.1)
Qwen-2.5 14B IT	37.6	61.6	84.2	75.2	48.8	47.2	44.0	40.4	21.8	66.0	80.8	55.2
trans-test	33.4	66.0	84.2	70.4	57.3	53.2	49.2	41.4	26.2	70.2	82.3	57.6 (+2.4)
LIDR (Yang et al., 2025b)	54.0	71.2	87.6	81.2	65.2	51.6	66.8	50.8	38.6	69.8	86.0	65.7 (+10.5)
S1 (Yong et al., 2025)	60.4	86.4	**90.4**	86.8	80.4	**80.8**	**89.2**	**54.0**	40.4	76.0	85.2	75.6 (+20.4)
self-cons (ours)	**65.6**	**89.2**	88.4	**87.8**	**82.2**	**80.8**	88.4	49.6	**44.6**	**81.6**	**88.0**	**76.9** (+21.7)

(a) **Qwen2.5 baselines**

.

	bn	de	en	es	fr	ja	ru	sw	th	zh	avg
Llama3 8B-IT	42.4	58.8	73.2	64.4	57.2	40.8	58.8	31.6	35.6	49.6	51.2
MAPO (She et al., 2024)†	38.4	50.8	71.6	62.8	59.2	38.4	62.0	30.8	47.2	54.0	51.5 (+0.3)
LIDR (Yang et al., 2025b)	47.2	57.0	72.0	64.0	60.8	47.6	62.0	32.8	**53.6**	59.2	55.6 (+4.4)
self-cons (ours)	**64.4**	**76.4**	**82.4**	**74.8**	**72.0**	**57.2**	**69.6**	**59.2**	41.6	**68.4**	**65.5** (+14.3)
Llama2-13B	35.2	44.4	53.2	48.0	48.4	43.2	47.6	42.8	46.8	48.8	45.8
MathOctopus-13B (Chen et al., 2024)†	42.0	49.2	51.6	53.2	49.6	39.6	47.6	46.0	46.0	51.2	47.6 (+1.8)
MetaMathOctopus-13B (She et al., 2024)†	41.6	**60.1**	66.8	61.1	60.8	57.3	59.1	50.9	52.1	53.1	51.5 (+5.7)
MAPO-13B (She et al., 2024)†	**44.8**	59.2	**71.6**	**62.8**	**63.6**	**56.0**	**59.3**	**55.2**	**47.6**	**59.6**	**58.0** (+12.2)
Llama2-7B	18.8	38.0	52.0	39.2	36.4	27.2	33.6	23.6	21.6	31.6	32.2
INCLINE-7B (Wang et al., 2025a)†	**48.4**	**51.2**	46.8	**45.2**	37.6	**44.8**	38.0	**38.8**	**43.2**	42.8	**43.7** (+11.5)
MathOctopus-7B (Chen et al., 2024)†	33.2	43.6	**54.8**	**45.2**	**38.0**	35.6	**48.4**	38.4	36.4	**45.2**	41.9 (+9.7)

(b) **Prior published methods.**

Table 1: **Main results on MGSM (accuracy).** Our method outperforms all Qwen2.5 baselines (a), with gains growing as model size increases, and exceeds prior published methods (b). All Qwen baselines are trained by us. We underline the overall best performance for each language.† indicates results reported by the method authors

aggregating multiple generations—e.g., via majority voting at high sample counts—substantially mitigates translation errors and yields reliable baseline performance (see App. B.2 for details).[4]

4 Experimental Setup

Base Model We conduct our experiments and ablations on the Qwen-2.5 family of models (Yang et al., 2024). We chose those models since they (i) are pre-trained and supervised fine-tuned on over 29 languages, (ii) come in a handful of practical sizes, facilitating ablations at scales, and (iii) widely used in multilingual research. In addition, we experiment with the Llama3 8B model (Grattafiori et al., 2024) to compare with prior work that use Llama-based models.

[4]Training on noisy signals is common in machine learning and can improve generalization. For example, although applying dropout at test time can hurt performance, it remains a standard and effective training technique.

Baselines We evaluate our approach against two categories of baselines.

(A) Prompt-based methods. We consider two prompting strategies: Original model performance using *native-cot*, where the model performs reasoning and produces answers in the original language of the query, and *translate-test*, where the query is first translated into English and the model subsequently solves the problem in English (Gao et al., 2025).

(B) Finetuning methods. We include two of the best-performing approaches that improve reasoning via finetuning. The first is test-time scaling (S1) (Yong et al., 2025), which applies SFT on English data to enhance reasoning capabilities. The second is LIDR (Yang et al., 2025b), which leverages DPO to improve multilingual reasoning performance. We train both methods on the Qwen2.5 models ourselves using the official code bases. In addition, we report relative performance gains compared to the prior published methods that use Llama

base models.

Monolingual Training Dataset Our method only requires English questions for training. We randomly sample 1k questions from the training split of the English-only GSM8k dataset (Cobbe et al., 2021). The final output is parsed using the standard regular expressions from the original benchmark. As discussed in §3, we use self-translation to obtain parallel questions, and analyze the impact of manual translation in §6.4

Evaluation Data Main evaluation and ablations are done on MGSM (Shi et al., 2022), which consists of 250 math word problems in 10 languages, which were sampled and manually translated from the GSM8K test set. In addition, we evaluate the generalization capabilities on 3 OOD datasets: mGPQA (Huang et al., 2025), MMath (Luo et al., 2025), and PolyMath (Wang et al., 2025b).

Training Environment and Hyperparameters We use GRPO as the RL algorithm and use VeRL (Sheng et al., 2024) for the training backend. We use a batch size of 32 parallel prompt sets (making a total of 320 prompts per step across all the 10 languages), with 32 rollouts per language and prompt, resulting in a total of 160 steps. The mini-batch size is set to 64 prompts (5 mini-batches per step), the KL coefficient to 0.001 and the learning rate to 1E-06. We set the temperature to 0.4.[5] All models are trained using 4×A100 GPUs. Other hyperparameters are left at the default values.

5 Results

Main Results Table 1 contains the main results of our method. As shown in Table 1a, we achieve average gains of 7.8, 15.1, and 21.7 points in Qwen2.5-IT 1.5B, 7B, and 14B, respectively, with improvements across nearly all languages. Notably, gains increase consistently with model size. We note that in Qwen-2.5 14B , Russian (ru) and Telugu (te) achieve the highest relevant performance gains: 102% and 104% respectively. We also observe degradation in Swahili for Qwen-2.5 7B and in English and Telugu in the Qwen-2.5 1.5B model.

Models trained with S1 test-time scaling remain competitive with our approach. In particular, they outperform our method on 3 languages for both Qwen-2.5 14B and Qwen-2.5 7B . However, these gains come at the cost of substantially longer reasoning traces, requiring approximately 3× more tokens. In contrast, our method achieves stronger overall performance with considerably more efficient token usage (see App. D.2). Furthermore, the improvements from S1 are driven primarily by the *quote-and-think* paradigm (Yong et al., 2025), where the model performs intermediate reasoning in English before generating the final answer in the target language. Consequently, these gains do not necessarily indicate improved intrinsic multilingual reasoning ability (see App. D.1).

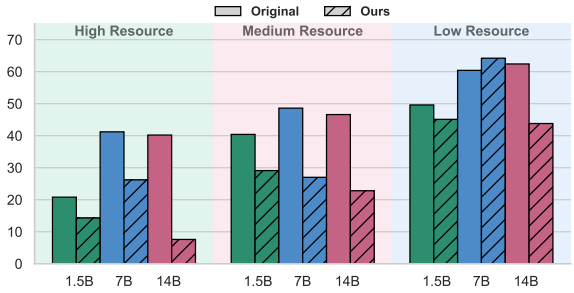

Figure 3: **Maximum performance difference relative to English within each resource tier on MGSM for Qwen2.5 models.** Our method significantly reduces the performance gap, with exception in the low-resource languages of the 7B model.

Performance Difference w.r.t English English is the best-performing language in the all baseline models. We examine the performance difference between English and other languages before and after training with our method. We partition languages into high-, medium-, and low-resource groups according to their representation in Common Crawl archives.[6] and show the maximum performance gap within each tier in Figure 3. Notably, the gap in performance is reduced up to 4 folds in High-resource languages for the Qwen-2.5 14B model. This suggests that cross-lingual performance gaps are not solely attributable to differences in factual or linguistic coverage, but may instead arise from inconsistencies in multilingual reasoning behavior. We observe similar behavior in the Medium-resource tier. Across model sizes, our method yields a comparable performance gap of 23-28% in the Medium-resource group. We observe that the difference increases in Low-resource languages for the Qwen-2.5 7B model, which can be explained by the degradation observed in Swahili

[5] Setting the temperature to a higher value sometimes led to unstable training. We leave investigating this for future work.

[6] commoncrawl.github.io/cc-crawl-statistics/plots/languages

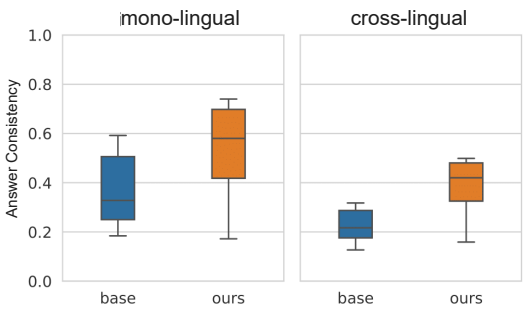

Figure 4: **Monolingual and Cross-lingual consistency** of Qwen-2.5 7B model on the MGSM benchmark. Our method improves both monolingual and cross-lingual consistency of the model.

for that particular model.

Although our method yields the largest gains for low-performing languages, they still lag behind high-resource languages. This reflects limitations of the multilingual backbone model (Qwen), whose weaker language proficiency in these languages constrains our method's efficiency.

Comparison to Prior Methods Head-to-head comparison with alternative methods to improve multilingual reasoning is obscured by different test sets, languages or base models. In Table 1b we report results on MGSM or other methods including INCLINE (Wang et al., 2025a), Mathoctopus (Chen et al., 2024), MetaMathOctopus (She et al., 2024), LIDR (Yang et al., 2025b) and MAPO (She et al., 2024).[7] The methods are applied to Llama2 or Llama3 base models. We run our self-consistency method on Llama3-8B model, and obtain the best results on average.

6 Analysis and Discussion

Below we analyze whether our method improves the cross-lingual self-consistency of the model, whether it generalizes to unseen domains and languages, the impact of the target distribution and that of multilingual datasets.

6.1 Cross-Lingual Self-Consistency

Monolingual self-consistency methods work by minimizing the entropy of the marginalized answer distribution (see §2). For easier interpretability, we next analyze the monolingual self-consistency of different models by taking the frequency of the most repeated answer instead:

$$\max_i \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\left(y_i^{(l)} = y_j^{(l)}\right) \tag{10}$$

Similarly, we measure cross-lingual self-consistency for each language l by taking the maximum frequency with which any of its answers appears across the rest of the languages:

$$\max_i \frac{1}{N(L-1)} \sum_{\substack{k=1 \\ k \neq l}}^{L} \sum_{j=1}^{N} \mathbf{1}\left(y_i^{(l)} = y_j^{(k)}\right) \tag{11}$$

We analyze self-consistency on MGSM before and after applying our method to Qwen-2.5 7B . To that end, we sample 32 parallel responses per example and language, and compute the average monolingual and cross-lingual self-consistency for each language. As shown in Figure 4, our approach greatly improves both consistency measures, with monolingual consistency remaining higher than cross-lingual consistency. In Appendix A we show that self-consistency improves as training progresses.

We also analyze the consistency of our method on the final accuracy performance using majority voting (Maj@k) with varying sample sizes. As shown in Figure 5, models trained with our approach reach near-peak performance at relatively low k: Maj@8 achieves over 95% of the accuracy achieved at Maj@32, suggesting our method is optimizing for inference compute. All model sizes exhibit the same behavior, as we show in Appendix B.1. We observe high inconsistency in the results of S1 trained models at higher values of K, mainly in the medium- and low-resource languages. We also note that prompt-based baselines — particularly self-translation— continue to benefit from increasing k, making it a reliable basis for our method.

6.2 Generalization

Out-of-distribution Generalization We evaluate our models trained with the GSM8k sample on three Out-of-distribution benchmarks: mGPQA, MMATH, and PolyMath. As shown in Table 2, our method achieves the best generalization performance. On mGPQA, we achieve an average gain of 2.1%, 4.3%, and 4.7% for the 1.5B, 7B, and 14B models, respectively, surpassing that of S1-trained models. On MMath, our method achieves up to **6.2%** gain in the 14B model. Conversely, LIDR has

[7]Note that Telugu was not reported for any prior model.

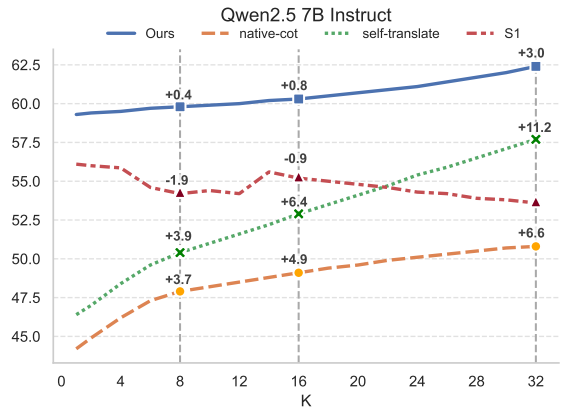

Figure 5: **Average accuracy using Majority@k** performance of the Qwen-2.5 7B IT model on the MGSM benchmark by taking the Maj@k for $k \in [1 \ldots 32]$. Vertical dashed-lines indicate the relative gain from Maj@1 for k={8,16,32} for each method.

Model Size	Method	mGPQA	MMATH	PolyMath
1.5B	**base**	22.8	46.9	7.5
	LIDR	18.8 (-4.0)	45.0 (-1.9)	5.9 (-1.6) †
	S1	24.5 (+1.7)	**53.8** (+6.9)	8.0 (+0.5)
	self-cons (ours)	**24.9** (+2.1)	48.8 (+1.8)	8.1 (+0.6)
7B	**base**	24.6	71.1	18.0
	LIDR	23.5 (-1.1)	67.0 (-3.9)	16.2 (-1.8) †
	S1	25.7 (+2.9)	**74.7** (+3.6)	20.2 (+2.2)
	self-cons (ours)	**28.9** (+4.3)	74.3 (+3.2)	20.5 (+2.5)
14B	**base**	28.2	72.2	17.5
	LIDR	29.4 (+0.8)	70.5 (-1.7)	15.2 (-2.3)†
	S1	32.1 (+3.9)	77.4 (+5.2)	19.5 (+2.0)
	self-cons (ours)	**32.9** (+4.7)	**78.4** (+6.2)	20.9 (+3.4)

Table 2: **Generalization of Qwen2.5-Instruct models on Out of Distribution (OOD) benchmarks**. Results on mGPQA, MMATH, and PolyMath (average of its 4 subtasks). Note that our models were trained on questions from GSM8K only. † indicates best run results

an adversarial impact on the performance, causing performance degradation on all OOD benchmarks. The PolyMath benchmark is the furthest from the GSM8k distribution, yet our method achieves the best relative performance gains. This corroborates the strength of our method in cross-domain generalization. We also observe that generalization benefits from model scale, indicating that larger models are more capable of leveraging our training signal.

Cross-lingual Generalization We train the Qwen-2.5 1.5B IT and Qwen-2.5 7B models on a random subset of 1, 2, and 3 languages and evaluate their performances on the remaining, unseen languages. As shown in Table 3, our method achieves substantial improvements in the unseen languages, **reaching up to 18.2% for the Qwen-2.5 7B -**

Model Size	L_{src}	Δ_{seen}	Δ_{unseen}
1.5B	bn	16.8	4.0
	bn,fr	13.3	4.2
	de,es,th	13.8	6.2
7B	bn	23.8	15.6
	bn,fr	15.7	18.2
	de,es,th	19.6	15.2

Table 3: **Generalization to Unseen Languages on MGSM.** Models trained on L_{src} languages are evaluated on held-out languages. Δ: difference in performance from the base model.

IT model. Notably, when the model is trained on a single source language—e.g., Bengali—the performance gain in that language is maximized. Nonetheless, including additional languages during training enhances the model's overall generalization to other unseen languages.

6.3 Different Target Distributions

We now analyze the strategies to estimate the target output distribution $q(y)$ used in reward modeling (see § 3). As shown in Table 8, using the source language outputs to construct the target distribution yields the least improvement. However, it can be useful in cases where the model is initially indecisive, such as Qwen-2.5 1.5B .

In contrast, constructing the target distribution from the outputs of English or combining the outputs of all languages achieves competitive performance. Specifically, using English outputs yields 1.4% more improvements than the combined outputs for Qwen-2.5 7B in MGSM. We hypothesize this is due to the less noisy English outputs and that English yields significantly higher base performance (at least 18% above other languages). Nonetheless, we believe that the choice of combined output is favorable, especially in domains and models where English performance is not dominant or other languages perform competitively. In addition, the fact that our method performs well even when considering weaker languages makes it robust for wide applicability.

6.4 Access to multilingual data for training

Our main approach accounts for domains where reasoning data is only available in one or a few languages by leveraging the translation capability of the model across supported languages. We now analyze the efficiency of our method in cases when multilingual data is accessible.

We post-train our models on the Multilingual

Model Size	Strategy	MGSM (avg)
1.5B	Source Lang	38.7 (+5.6)
	English	40.4 (+7.3)
	Combined	**40.9** (+7.8)
7B	Source Lang	48.8 (+4.6)
	English	**61.0** (+16.8)
	Combined	59.6 (+15.4)

Table 4: **Performance of Qwen2.5-IT models trained with our method using different target distribution estimation strategies.** All models are trained on the monolingual data and evaluated on the MGSM benchmark.

Grad School Mathematics (MGSM) dataset (Shi et al., 2022). The dataset includes 250 examples from the test set of the GSM8k benchmark, manually translated into 10 languages besides English. We train on 25% of the MGSM sampled uniformly over languages and evaluate on the remaining 75%. As shown in Table 5, this setup improves the results for the smaller models, but gets comparable average results for the largest model. We hypothesize that the largest model has enough quality to self-translate, and that, after all, it is the self-consistency of the trajectories that matters most, and that adequate translation of the questions suffices.

7 Related Work

Multilingual Reasoning with Supervised Fine-Tuning. Prior work addresses cross-lingual reasoning gaps by fine-tuning models on translated question pairs to improve alignment across languages (Zhu et al., 2024; Chen et al., 2024). However, the scarcity of high-quality parallel data and the noise introduced by machine translation limit scalability and performance on complex reasoning tasks, making English-only reasoning (English-cot) remain the most effective baseline (Liu et al., 2025; Ranaldi et al., 2025). Some works add a second step with preference optimization, which requires external translation, and use English trajectories as preferred with respect to native trajectories (She et al., 2024; Yang et al., 2025b)

Multilingual Reasoning with Inference-Time Intervention. To bypass data limitations, recent studies explore inference-time strategies adapted from monolingual reasoning. Longer chains of thought (Yong et al., 2025) and multilingual ensembling via majority voting (Gao et al., 2025) both improve cross-lingual performance, with English often serving as the most reliable reasoning language. As an alternative (Huang et al., 2024; Wang et al., 2025a), response generation can be manipulated using the logits to mimic those of English. Nevertheless, they perform poorly compared to supervised fine-tuning methods.

Model Size	Train Dataset	Acc. (avg)
1.5B	GSM8k	40.9 (+7.8)
	MGSM	43.5 (+10.4)
7B	GSM8k	59.3 (+15.1)
	MGSM	61.2 (+17.0)
14B	GSM8k	76.5 (+20.3)
	MGSM	76.5 (+20.3)

Table 5: **Impact of using manual translations in training Qwen2.5-IT models.** Results when training on a sample of the Multilingual data (MGSM, manual translations) compared to training on (GSM8K, self-translation before training).

Monolingual Reasoning. Outcome-based reward modeling has recently shown that enforcing output consistency can improve reasoning without ground-truth supervision, but existing work is confined to English (Agarwal et al., 2025; Prabhudesai et al., 2025; Elhady et al., 2025; He et al., 2026). We extend this paradigm to multilingual reasoning.

8 Conclusion and Future Work

In this paper, we show that incentivizing cross-lingual consistency improves reasoning in the source language. Our method uses only monolingual questions, applying reinforcement learning with an unsupervised cross-lingual consistency loss to further train an existing model. It assumes no preferred language and enforces consistency across all languages. Experiments on MGSM show substantial gains, especially for weaker languages, while also improving stronger ones and reducing performance gaps. Results on OOD benchmarks indicate good generalization, though with smaller gains. Since our method relies only on monolingual questions, we hypothesize that training on more diverse domains could further enhance generalization.

In the future we would like to explore further domains, and extend our method to non-verifiable tasks. All things considered, our findings pave the way to improve the multilingual reasoning performance at large without any extra data annotation.

Limitations

Our experiments are limited to sub 20B LLMs, primarily due to compute constraints. While we observe that our method is increasingly effective as the size grows it would be interesting to explore its behaviour for larger models. In addition, we used the Qwen2.5 Models; It would be efficient to see the impact on more recent models such as Qwen3 (Yang et al., 2025a).

Our training and evaluation were highly constrained by the data. The limited availability of reliable multilingual reasoning benchmarks—particularly for low-resource languages—restricts broader validation and remains a hindering challenge. We hope to address some of these limitations in future work.

Acknowledgments

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *Preprint*, arXiv:2505.15134.

Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Self-questioning language models. *Preprint*, arXiv:2508.03682.

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, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.

Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. Emergent abilities of large language models under continued pretraining for language adaptation. *Preprint*, arXiv:2506.00288.

Changjiang Gao, Xu Huang, Wenhao Zhu, Shujian Huang, Lei Li, and Fei Yuan. 2025. Could thinking multilingually empower llm reasoning? *Preprint*, arXiv:2504.11833.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The multilingual mind : A survey of multilingual reasoning in language models. *Preprint*, arXiv:2502.09457.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangziqi Zhao, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, Xiusi Chen, Youbang Sun, Xingtai Lv, Xuekai Zhu, Li Sheng, Ran Li, Huan ang Gao, Yuchen Zhang, Bowen Zhou, and 2 others. 2026. How far can unsupervised rlvr scale llm training? *Preprint*, arXiv:2603.08660.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficient boosting llm reasoning in non-english languages. *Preprint*, arXiv:2405.17386.

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. 2025. Cross-lingual optimization for language transfer in large language models. *Preprint*, arXiv:2505.14297.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. When natural language is not enough: The

limits of in-context learning demonstrations in multilingual reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7369–7396, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

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)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.

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. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025a. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report.

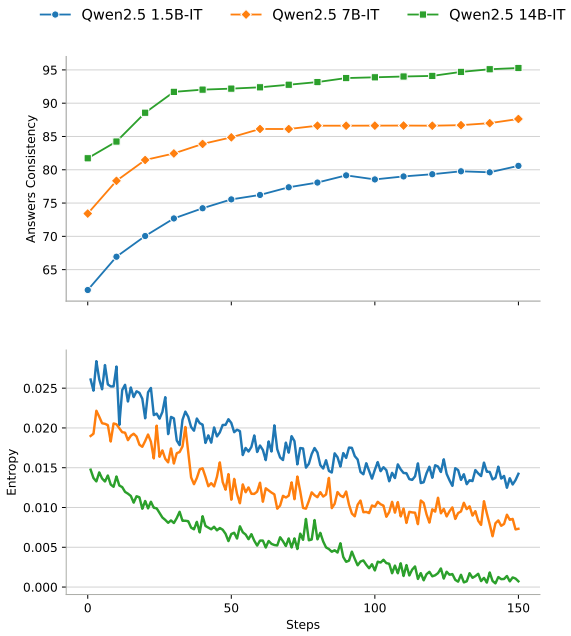

Figure 6: **Training curves. Top:** Cross-lingual consistency improvement throughout training. **Bottom:** Entropy decreases throughout training.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025b. Language imbalance driven rewarding for multilingual self-improving. In *The Thirteenth International Conference on Learning Representations*.

Zheng-Xin Yong, M. Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H. Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *Preprint*, arXiv:2505.05408.

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*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

A Self-consistency Training Curves

In Figure 6 (top), we show that our method yields a steady progression for all models. The progression improves steadily for all models, and the upward trend suggests further improvements could be achieved by scaling up the training compute. Concurrent to the improvements in cross-lingual consistency, we show in Figure 6 (bottom) that the entropy of the model's policy is steadily regularized as training progresses. This supports our view of the method as a cross-entropy regularization technique and links the improvement in cross-lingual

consistency with the improvement of the model entropy.

B Self-consistency Accuracy Performances

B.1 Majority@k Performances

Figure 7 shows the performance of Qwen-2.5 IT models using Majority@k, for $k \in \{1, 8, 32\}$. Our method achieves near-peak performance at lower samples, making it more efficient with respect to inference budget.

B.2 Detailed Performances

In Table 6, we show the performance of different methods by computing the accuracy with majority vote at different sample sizes (Maj@k). Ideally, a self-consistent model achieves the best performance at a lower sample size, hence a more efficient inference token consumption. Models trained with our approach achieve their highest performance at lower K values, with Maj@8 achieving more than 95% of that of Maj@32.

In addition, results show that the performance of prompt-based methods, specifically self-translation, improves consistently with higher K, corroborating our assumption that they provide a reliable basis for our method.

C Detailed Performances

Tables 2 – 9 show the detailed performances of our method training.

D Comparison to S1 trained models

Models trained with S1 test-time scaling in English exhibit impressive improvements in other languages (Yong et al., 2025). They achieve competetive performances to our method. We analyze the reasoning traces of the S1 models and ours to better compare the methods.

D.1 Reasoning Traces

In Box 1, we show traces of Qwen2.5 14B trained with our method and the s1-method on a Spanish question. Models trained with S1 test-time scaling technique exhibit a quote-and-think paradigm where they do reasoning first in English and deliver the final response –and occasionally short chain of thought– in the prompt language. Albeit efficient, this does not imply that the S1 training recipe induces better multilingual reasoning, but rather surfaces the model's translation capabilities as a proxy for enhanced reasoning.

D.2 Test-time Token Consumption

S1 models tend to do longer reasoning in English, while our method arrives at the same answer with shorter and language-consistent chain-of-thought. In Table 10, we show the average and maximum token length of the responses per language using Qwen-2.5 14B trained with our method and with S1 test-time scaling.

Figure 7: **Accuracy Majority@k** performance of the Qwen-2.5 IT models on the MGSM benchmark by taking the Maj@k for $k \in [1 \ldots 32]$

Model	Method	K	bn	de	en	es	fr	ja	ru	sw	te	th	zh	avg
Qwen2.5-1.5B-Instruct	native-cot	1	12.0	41.2	52.4	48.8	48.1	31.6	46.4	2.8	4.4	34.6	41.2	33.0
		8	12.0	45.2	55.6	48.4	54.0	37.2	44.8	2.8	4.0	31.6	50.8	35.1
		32	16.4	48.4	58.4	49.2	56.0	37	49.8	3.2	6.2	33.6	56.2	37.7
	self-translate	1	14.4	44.2	52.4	50.0	42.1	23.6	38.8	8.4	12.4	36.8	40.8	33.1 (+0.1)
		8	14.2	48.0	56.8	49.0	54.8	36.8	48.4	3.2	6.0	32.8	56.0	36.9 (+1.8)
		32	8.0	45.2	56.8	48.4	56.0	37.2	48.0	8.2	12.0	32.8	56.0	37.1 (+4.1)
	eng-cot	1	13.2	41.8	52.4	51.6	50.6	27.6	46.8	3.2	2.0	36.6	52.8	34.4 (+1.4)
		8	14.2	48	54.8	49.0	52.0	30.8	48.4	3.2	6.0	32.8	54.6	35.8 (+0.7)
		32	16.4	48.8	56.8	49.2	54.8	36.8	49.4	3.4	12.8	33.6	57.6	38.1 (+5.1)
	ours	1	21.2	53.6	50.3	63.2	56.4	36.0	55.2	5.2	7.2	48.0	54.0	40.9 (+7.9)
		8	24.0	56.4	56.0	66.0	57.2	39.0	57.6	4.8	7.2	48.0	58.4	43.1 (+8.0)
		32	24.4	56.8	56.4	67.2	57.2	39.8	60.2	5.2	7.2	48.4	58.4	43.7 (+10.7)
Qwen2.5-7B-Instruct	native-cot	1	30.0	60.4	78.6	40.2	48.2	49.8	37.4	19.8	18.2	54.8	49.2	44.2
		8	30.8	62.4	80.6	57.2	51.6	50.4	44.2	21.6	20.0	53.4	54.6	47.9
		32	31.2	63.2	88.4	58.6	55.2	52.8	48.0	23.6	21.2	56.2	60.8	50.8
	self-translate	1	33.4	67.6	78.6	50.3	54.4	43.6	39.2	11.4	18.6	60.2	52.6	46.4 (+2.2)
		8	34.3	69.8	80.6	71.6	58.2	44.1	46.3	12.4	20.4	58.7	58.4	50.4 (+2.5)
		32	30.2	72.8	88.4	77.2	71.6	61.2	68.0	24.4	20.0	51.2	69.6	57.7 (+13.5)
	eng-cot	1	34.2	59.2	80.0	65.0	70.5	41.8	64.8	17.4	11.2	44.0	60.9	49.9 (+5.7)
		8	35.1	61.2	82	62.0	70.6	42.3	76.6	19.0	12.3	42.9	67.6	52.0 (+4.1)
		32	31.8	73.2	88.4	77.6	70.0	62.8	67.2	23.6	12.0	52.4	68.8	57.1 (+12.9)
	ours	1	54.8	75.2	81.8	62.8	71.6	55.6	76.0	17.6	24.0	73.2	59.6	59.3 (+15.1)
		8	56.3	77.7	80.4	62.0	70.2	56.3	76.2	19.2	22.0	71.3	66.1	59.8 (+11.9)
		32	58.2	78.0	82.6	57.0	74.6	60.8	78.8	24.6	24.8	78.4	68.6	62.4 (+18.2)
Qwen2.5-14B-Instruct	native-cot	1	37.6	61.6	84.2	75.2	48.8	47.2	44.0	40.4	21.8	66.0	80.8	55.2
		8	38.6	63.6	86.3	76.4	52.2	47.8	46.0	44.1	24.0	64.3	82.0	56.8
		32	40.2	64.6	94.8	80.2	56.8	52.0	56.4	48.2	28.4	68.4	84.0	61.3
	self-translate	1	33.4	66.0	84.2	70.4	57.3	53.2	49.2	41.4	26.2	70.2	82.3	57.6 (+2.4)
		8	34.3	68.1	80.6	70.0	61.3	63.8	72.6	42.8	22.8	68.5	81.2	60.5 (+3.7)
		32	28.2	80.0	80.0	72.4	81.6	70.0	86.2	44.0	22.0	66.8	84.0	65.0 (+9.8)
	eng-cot	1	56.6	83.2	83.8	72.0	77.6	39.2	73.6	48.8	25.6	73.4	83.0	65.2 (+10.0)
		8	58.1	86.0	83.0	68.7	77.7	64.2	76.0	50.2	28.1	71.6	92.1	68.7 (+11.9)
		32	47.6	84.2	84.0	73.6	81.2	71.6	84.0	54.4	36.4	77.2	78.4	70.2 (+15.0)
	ours	1	65.6	89.2	88.4	87.8	82.2	80.8	88.4	49.6	44.6	81.6	88.0	76.9 (+21.7)
		8	67.4	92.2	86.9	86.7	80.6	81.8	90.6	54.2	48.0	82.6	88.4	78.1 (+21.3)
		32	66.2	90.0	89.2	90.4	84.0	82.6	91.6	54.2	49.8	84.6	92.4	79.5 (+24.3)

Table 6: **Accuracy for Majority@K for base and models trained with our method on the MGSM benchmark.**

	Method	bn	de	en	es	fr	ja	ru	sw	te	th	zh	avg
Qwen-2.5 1.5B IT	base	17.0	20.1	**28.5**	21.1	23.1	20.1	**27.2**	2.2	5.9	20.3	25.7	22.8
	english-cot	**22.3**	**25.0**	28.5	23.7	21.0	**23.0**	18.5	**12.9**	**13.7**	21.7	22.1	21.2 (-1.6)
	self-cons (ours)	21.8	22.7	26.9	**26.3**	**25.7**	21.7	26.8	2.3	6.2	**24.1**	**27.6**	**24.9** (+2.1)
Qwen-2.5 7B IT	base	22.8	25.7	30.4	27.7	33.3	24.3	26.3	8.3	16.5	23.4	32.1	24.6
	english-cot	26.1	**31.9**	30.4	32.1	27.9	**32.6**	**29.0**	**16.2**	**27.2**	26.1	32.4	28.4 (+3.8)
	self-cons (ours)	**26.7**	31.0	**32.4**	**32.5**	**39.9**	**27.7**	28.8	12.2	19.6	**29.2**	**37.7**	**28.9** (+4.3)
Qwen-2.5 14B IT	base	23.4	35.7	32.8	31.9	32.1	27.2	33.3	15.0	17.6	27.5	34.2	28.2
	english-cot	**34.8**	37.5	32.8	**37.9**	36.7	29.7	32.0	19.9	19.6	**32.9**	**40.4**	32.2 (+4.0)
	self-cons (ours)	29.5	**39.1**	**36.6**	33.5	**37.3**	**32.1**	**34.6**	**23.9**	**25.7**	31.5	38.4	**32.9** (+4.7)

Table 7: **Generalization to other domains**. Results on mGPQA for our models (which were trained on questions from GSM8K).

	Strategy	bn	de	en	es	fr	ja	ru	sw	te	th	zh	avg
Qwen-2.5 1.5B IT	**Source Lang**	30.8	48.4	55.2	52.4	53.6	35.6	52.0	4.0	4.8	40.8	48.8	38.7 (+5.6)
	English	**38.8**	47.6	**57.8**	53.6	**59.6**	32.0	50.4	4.4	7.2	44.0	48.8	40.4 (+7.3)
	Combined	21.2	**53.6**	50.3	**63.2**	56.4	**36.0**	**55.2**	**5.2**	**7.2**	**48.0**	**54.0**	**40.9** (+7.8)
Qwen-2.5 7B IT	**Source Lang**	43.2	68.8	66.0	52.0	58.4	50.2	48.4	17.2	22.6	57.6	52.8	48.8 (+4.6)
	English	54.0	**79.6**	80.8	62.6	**74.0**	**60.0**	69.6	**20.0**	**34.4**	**75.2**	**60.8**	**61.0** (+16.8)
	Combined	**54.8**	75.2	**81.8**	**62.8**	71.6	55.6	**76.0**	17.6	24.0	73.2	59.6	59.6 (+15.4)

Table 8: **Performance of different target distribution estimation strategies.** All models are trained on the monolingual data and evaluated on the MGSM benchmark.

	Train Dataset	bn	de	en	es	fr	ja	ru	sw	te	th	zh	avg
Qwen-2.5 1.5B IT	**GSM8k**	21.2	53.6	50.3	63.2	56.4	36.0	55.2	5.2	7.2	48.0	54.0	40.9 (+7.8)
	MGSM	29.2	53.2	56.8	63.0	56.8	41.6	55.2	5.6	6.4	50.0	61.2	43.5 (+10.4)
Qwen-2.5 7B IT	**GSM8k**	54.8	75.2	81.8	62.8	71.6	55.6	76.0	17.6	24.0	73.2	59.6	59.3 (+15.1)
	MGSM	63.2	78.8	78.5	72.0	64.4	62.4	68.0	21.2	32.8	75.6	62.8	61.2 (+17.0)
Qwen-2.5 14B IT	**GSM8k**	65.6	89.2	88.4	87.8	82.2	80.8	88.4	49.6	44.6	81.6	88.0	76.9 (+21.7)
	MGSM	62.0	86.4	88.0	90.4	85.2	82.8	88.2	47.6	41.6	83.6	85.2	76.5 (+20.3)

Table 9: **Impact of using manual translations in training.** Results when training on a sample of the Multilingual data (MGSM, manual translations) compared to training on (GSM8K, self-translation before training). Both settings are evaluated on the questions of the holdout sample of the MGSM dataset (75%).

Box 1. Comparing the reasoning trace of the s1's *quote-and-think* vs ours

Question: Los patos de Janet ponen 16 huevos por día. Ella come tres en el desayuno todas las mañanas y usa cuatro para hornear magdalenas para sus amigos todos los días. Vende lo que sobra en el mercado de productores diariamente a $2 el huevo fresco de pato. ¿Cuánto gana en dólares todos los días en el mercado de productores?
(Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and uses four to bake muffins for her friends every day. She sells the leftover eggs at the farmers market daily for $2 per fresh duck egg. How much does she earn in dollars each day at the farmers market?)

s1's trace:
Okay, let me try to figure this out. So, Janet has some ducks that lay eggs. Let's see, they lay 16 eggs per day. Then, Janet uses some of them herself and bakes with some, and sells the rest at the market. The question is how much she earns each day from selling the leftover eggs. Each egg sells for $2 ... Wait, that seems straightforward ... so I think $18 is the correct answer. Respuesta: Janet's patos ponen 16 huevos diariamente. Ella consume 3 huevos para el desayuno y utiliza 4 huevos para hornear magdalenas, totalizando 7 huevos usados diariamente. Por lo tanto, los huevos sobrantes para vender son 16 - 7 = 9 huevos. Vendiendo cada huevo a $2, Janet gana 9 x $2 = $18 diariamente en el mercado de productores. **Respuesta Final \boxed{18}**

Self-consistency (ours) trace:
Primero, Janet tiene 16 huevos por día. Come 3 huevos y usa 4 huevos, por lo que en total usa 3 + 4 = 7 huevos. Esto deja 16 - 7 = 9 huevos para vender. Si vende cada huevo a $2, entonces gana 9 * 2 = $18 todos los días en el mercado de productores.**Respuesta: $18**

	bn	de	en	es	fr	ja	ru	sw	te	th	zh	avg
Qwen-2.5 14B IT	229	189	127	139	146	157	122	142	314	212	197	188.3
Ours	241	188	135	155	246	151	173	152	251	198	167	216.1
S1	430	358	270	440	296	356	320	340	714	410	290	366.1

Table 10: **Average response token length on the MGSM dataset.** Our models surpass the performance of S1 at much lower token length.