CROSSLINGUAL ON-POLICY SELF-DISTILLATION FOR MULTILINGUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have achieved remarkable progress in mathematical reasoning, but this ability is not equally accessible across languages. Especially low-resource languages exhibit much lower performance. To address this, we propose ***C**rosslingual **O**n-**P**olicy **S**elf-**D**istillation* (**COPSD**), which transfers a model's own high-resource reasoning behavior to low-resource languages. COPSD uses the same model as student and teacher: the student sees only the low-resource problem, while the teacher receives privileged crosslingual context, including the problem translation and reference solution in English. Training minimizes full-distribution token-level divergence on the student's own rollouts, providing dense supervision while avoiding the sparsity and instability of outcome-only reinforcement learning (RL). Experiments on 17 low-resource African languages show that COPSD consistently improves mathematical reasoning across model sizes and outperforms alternative post-training baselines, including Group Relative Policy Optimization (GRPO) and supervised fine-tuning (SFT). Further analysis shows that COPSD improves answer-format adherence, benefits from larger inference-time generation budgets, mitigates repetitive degeneration, and generalizes to harder multilingual reasoning benchmarks, with especially large gains for lower-resource languages.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable progress in mathematical reasoning (Ahn et al., 2024; Yang et al., 2025a; Guo et al., 2025). A key driver of this progress is their ability to generate step-by-step reasoning traces, which can elicit strong problem-solving behavior (Wei et al., 2022). However, this capability remains far from multilingual. Models often struggle when reasoning in underrepresented languages (Hwang et al., 2025; Yong et al., 2025; Ghosh et al., 2025), which receive limited exposure during the pretraining stage and are rarely represented in high-quality reasoning supervision during the post-training stage (Qin et al., 2024; Yang et al., 2025b). As a result, a model may possess the latent capability to solve a problem, yet fail to access that capability when the problem and reasoning traces are expressed in a low-resource language.

A natural approach to this issue is to construct reasoning supervision directly in low-resource languages, e.g., by translating English reasoning traces into the target low-resource languages and then directly performing supervised fine-tuning (SFT) on these translated trajectories (Wu et al., 2025; Barua et al., 2026). Yet this approach faces several limitations. Machine translation can introduce noise and is prone to inconsistencies or errors in mathematical expressions, quantities, and logical dependencies (Petersen et al., 2023; Zhang et al., 2024). Moreover, translated reasoning traces may not match the model's own reasoning behavior and therefore can suffer from train-inference distribution mismatch (Agarwal et al., 2024; Gu et al., 2024). Another possibility is to use reinforcement learning (RL) with outcome-based rewards, where the model is rewarded when its final answer matches the ground truth (Schulman et al., 2017; Shao et al., 2024). However, such rewards can become extremely sparse in low-resource-language settings: if the model rarely produces correct answers, then binary outcome feedback provides little information about how intermediate reasoning should be improved, making RL sample-inefficient and potentially unstable (Lightman et al., 2024). These limitations suggest the need for a training signal that is both *dense* and *scalable*, while remaining aligned with the reasoning trajectories the model actually produces in low-resource languages.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this, we extend *on-policy self-distillation*, where a single model serves as both student and teacher policies under different contexts and learns from dense feedback on its own trajectories (Zhao et al., 2026b; Zhang et al., 2026b; Sang et al., 2026). We propose ***C**rosslingual **O**n-**P**olicy **S**elf-**D**istillation* (**COPSD**), which transfers English-accessible reasoning behavior to low-resource languages. In COPSD, the student observes only the low-resource problem, while the teacher additionally receives privileged crosslingual context, including the English problem and reference solution. The student first generates its own reasoning trajectory, and training minimizes a full-distribution token-level divergence between the student and teacher policies along this trajectory. This provides dense supervision at every decoding step while keeping training aligned with the reasoning paths the student policy actually explores.

We train Qwen3 models at three scales (1.7B, 4B, and 8B) with COPSD on 17 low-resource African languages and evaluate them on AfriMGSM (Adelani et al., 2025). Our results show that COPSD consistently improves over the base models and substantially outperforms GRPO and SFT . Further analysis shows that COPSD converges rapidly, improves answer-format adherence, and enables models to better leverage larger test-time generation budgets. We also evaluate COPSD on 8 languages from the more challenging PolyMath benchmark (Wang et al., 2025c), finding that its gains generalize beyond AfriMGSM and are especially pronounced for lower-resource languages.

Our contributions are summarized as follows: (**i**) We propose COPSD, a crosslingual on-policy self-distillation framework that uses high-resource language context as privileged information to improve low-resource reasoning. (**ii**) We demonstrate consistent improvements over base models and substantial gains over GRPO and SFT baselines across 17 low-resource African languages and multiple model sizes. (**iii**) We analyze training dynamics, answer-format adherence, and test-time scaling, showing that COPSD improves both accuracy and the effectiveness of low-resource reasoning trajectories. (**iv**) We show that COPSD generalizes to harder multilingual reasoning settings, with especially strong gains for lower-resource languages. (**v**) We will release our code and data to support future research on multilingual reasoning in low-resource languages.

2 Related Work

On-Policy Distillation. On-policy distillation (OPD) (Gu et al., 2024; Agarwal et al., 2024; Lu & Lab, 2025; Yang et al., 2026) has emerged as an effective alternative to both SFT (Yang et al., 2024; Chung et al., 2024; Ye et al., 2025) and outcome-based RL for improving LLM reasoning (Shao et al., 2024; Liu et al., 2025; Wen et al., 2025). Compared to SFT and RL, OPD combines on-policy supervision from student-generated trajectories with dense token-level teacher feedback, thereby reducing train-inference distribution mismatch while avoiding sparse sequence-level rewards (Agarwal et al., 2024; Gu et al., 2024; Zhao et al., 2026b). Recent work shows that effective OPD requires compatible teacher-student thinking patterns, as mismatches can hinder reasoning capability transfer (Li et al., 2026). This motivates *on-policy self-distillation*, where a single model serves as both student and teacher under different contexts to improve reasoning behavior (Zhao et al., 2026b; Zhang et al., 2026b; Kim et al., 2026; Sang et al., 2026). In recent work, Qin et al. (2026) propose multilingual self-distillation, with both on-policy and off-policy variants, to transfer safety alignment from high-resource to low-resource languages. Our work shares a similar OPSD mechanism but focuses on multilingual mathematical reasoning. Specifically, we use privileged teacher context to transfer the model's English-accessible reasoning behavior to low-resource languages.

Multilingual Reasoning. Multilingual reasoning concerns the ability of language models to solve reasoning problems consistently across languages, rather than relying primarily on English or other high-resource languages (Ghosh et al., 2025). Prior work shows that LLMs exhibit substantial crosslingual performance gaps (Tam et al., 2025; Zhao et al., 2026a; Liu et al., 2026; Ki et al., 2026; Zhang et al., 2026a), especially in low-resource languages, and may generate inconsistent or language-mixed reasoning traces (Qi et al., 2025; Wang et al., 2025a). To address these issues, existing methods often use *translate-and-test* pipelines (Qin et al., 2023; Huang et al., 2023; Zhu et al., 2024; Kang et al., 2026), *supervised fine-tuning* (Zhao et al., 2024; Zhang et al., 2024; Üstün et al., 2024; Lai & Nissim, 2024), *self-training* (Ranaldi & Pucci, 2025; Sutawika et al., 2026), and *reinforcement learning* (She et al., 2024; Ranaldi & Pucci, 2025; Wang et al., 2025b; Huang et al., 2025; Faisal et al., 2025; Zhang et al., 2026c). However, these approaches typically require translated reasoning rationales or sparse outcome rewards. In contrast, our proposed method COPSD

improves low-resource reasoning by using high-resource language context as privileged information and distilling dense token-level supervision from the same model on its own low-resource reasoning.

3 PRELIMINARY: ON-POLICY SELF-DISTILLATION

3.1 TEACHER AND STUDENT POLICIES

On-Policy Self-Distillation (OPSD) is a framework for improving reasoning without requiring a separate teacher model (Zhao et al., 2026b; Zhang et al., 2026b). Instead of distilling knowledge from an external model (Agarwal et al., 2024; Lu & Lab, 2025), OPSD instantiates the same model as both a student and a teacher under different conditioning contexts. Given a reasoning dataset $\mathcal{D} = \{(x, c)\}$, where x is a problem and c is privileged information such as a reference solution, OPSD defines two policies from the same model p_θ:

$$p_S(\cdot \mid x) \triangleq p_\theta(\cdot \mid x),$$
$$p_T(\cdot \mid x, c) \triangleq p_\theta(\cdot \mid x, c).$$

The student policy p_S observes only the problem, matching the inference-time setting, while the teacher policy p_T additionally conditions on privileged information. Although both policies share the same parameters, the teacher distribution is expected to provide a stronger learning signal because the reference solution helps the teacher policy favor continuations consistent with correct reasoning.

3.2 ON-POLICY TRAJECTORY SAMPLING

OPSD preserves the on-policy training paradigm by sampling trajectories from the **student** rather than from the teacher. For a problem x, the student generates a response

$$\hat{y} = (\hat{y}_1, \ldots, \hat{y}_N) \sim p_S(\cdot \mid x).$$

Both the student and teacher then evaluate this student-generated trajectory $\hat{y}$. At each decoding step n, they produce next-token distributions conditioned on the same prefix $\hat{y}_{<n}$:

$$p_S^n \triangleq p_S(\cdot \mid x, \hat{y}_{<n}),$$
$$p_T^n \triangleq p_T(\cdot \mid x, c, \hat{y}_{<n}).$$

3.3 DISTILLATION OBJECTIVE

The training objective minimizes the trajectory-averaged token-level divergence between the teacher and student distributions:

$$D(p_T \parallel p_S)(\hat{y} \mid x) = \frac{1}{N} \sum_{n=1}^{N} D(p_T^n \parallel p_S^n),$$

where D can be instantiated as a distributional divergence such as KL divergence (Kullback & Leibler, 1951). The overall OPSD objective is

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{(x,c)\sim\mathcal{D}}\, \mathbb{E}_{\hat{y}\sim p_S(\cdot|x)} \left[D(p_T \parallel p_S)(\hat{y} \mid x) \right].$$

Gradients flow only through the student policy, while the teacher serves as a fixed distributional target conditioned on privileged information.

4 METHODOLOGY

We introduce ***C**rosslingual **O**n-**P**olicy **S**elf-**D**istillation* (**COPSD**), which extends OPSD to multilingual reasoning. The key idea is to leverage high-resource language information as *privileged context*. In this work, we use English as the high-resource language, reflecting the English-centric nature of common LLM post-training (Shaham et al., 2024; Dang et al., 2024). During COPSD training, the student must reason from the low-resource problem alone, while the teacher is given additional privileged context (English problem and the English reference solution) that helps elicit a stronger reasoning distribution from the same model, as shown in Figure 1. This allows the model to transfer its own English-accessible reasoning behavior to low-resource languages without relying on an external teacher or (translated) target-language reasoning traces.

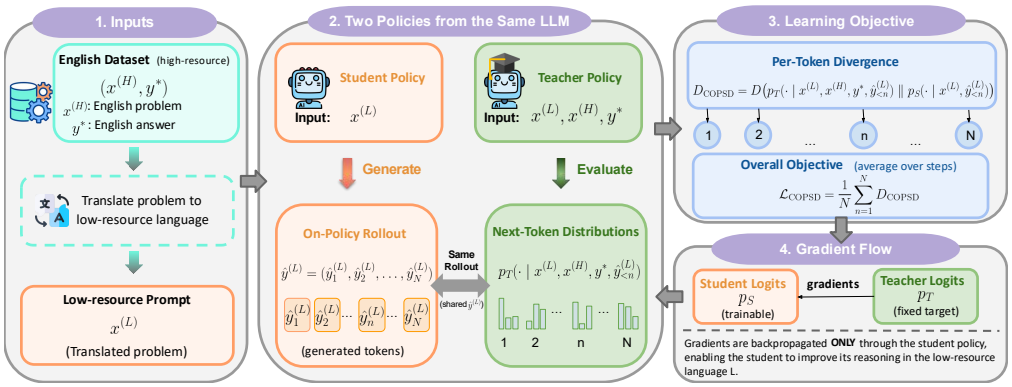


Figure 1: Overview of COPSD. Each problem is translated into a low-resource language as the student's input. The same LLM acts as both student and teacher: the student generates an on-policy rollout, while the teacher evaluates it with privileged English context and the reference solution. By minimizing per-token divergence along the rollout, COPSD transfers English-accessible reasoning behavior to improve reasoning in low-resource languages.

4.1 Crosslingual Learning Setup

We consider a multilingual reasoning dataset

$$\mathcal{D} = \{(x^{(L)}, x^{(H)}, y^*)\},$$

where $x^{(L)}$ denotes a problem in a low-resource language, $x^{(H)}$ denotes its English counterpart, and y^* denotes the full reference solution in English, including the reasoning steps and final answer.

Following OPSD, COPSD instantiates two policies from the same language model p_θ. The student policy observes only the low-resource problem:

$$p_S(\cdot \mid x^{(L)}) \triangleq p_\theta(\cdot \mid x^{(L)}).$$

The teacher policy additionally receives privileged high-resource context $c \triangleq (x^{(H)}, y^*)$, comprising the English problem and its reference solution:

$$p_T(\cdot \mid x^{(L)}, c) \triangleq p_\theta(\cdot \mid x^{(L)}, x^{(H)}, y^*).$$

Thus, the student matches the inference-time condition, while the teacher has access to information that can induce more reliable reasoning behavior.[1]

4.2 On-Policy Crosslingual Distillation

Given a problem $x^{(L)}$, the student generates a reasoning trajectory in the target low-resource language:

$$\hat{y}^{(L)} = (\hat{y}_1^{(L)}, \ldots, \hat{y}_N^{(L)}) \sim p_S(\cdot \mid x^{(L)}).$$

Both policies then evaluate the same student-generated prefix. At each step n, we have

$$p_S^n \triangleq p_S(\cdot \mid x^{(L)}, \hat{y}_{<n}^{(L)}),$$
$$p_T^n \triangleq p_T(\cdot \mid x^{(L)}, x^{(H)}, y^*, \hat{y}_{<n}^{(L)}).$$

COPSD then minimizes the token-level divergence between the teacher and student distributions along the student's own rollout:

$$D(p_T \,\|\, p_S)(\hat{y}^{(L)} \mid x^{(L)}) = \frac{1}{N}\sum_{n=1}^{N} D(p_T^n \,\|\, p_S^n),$$

where D is a distributional divergence, e.g., KL divergence. The training objective is formulated as

$$\mathcal{L}_{\text{COPSD}}(\theta) = \mathbb{E}_{(x^{(L)}, x^{(H)}, y^*)\sim\mathcal{D}} \mathbb{E}_{\hat{y}^{(L)}\sim p_S(\cdot \mid x^{(L)})} \left[D(p_T \,\|\, p_S)(\hat{y}^{(L)} \mid x^{(L)}) \right].$$

Gradients are backpropagated only through the student policy, enabling the student to improve its reasoning in the low-resource language L.

[1] During training, we control the explicit reasoning language of both the student and teacher policies to match the low-resource language of the student input; see §5.2.

5 EXPERIMENTS

5.1 MODELS

We conduct experiments with Qwen3 models (Yang et al., 2025a) of three sizes: `Qwen3-1.7B`, `Qwen3-4B`, and `Qwen3-8B` in our main experiments. Qwen3 models are pretrained on multilingual corpora and further post-trained on data dominated by high-resource languages such as English.

5.2 CONTROLLING THE REASONING LANGUAGE

LLMs may switch to English in their reasoning traces, even when prompted in a different target language (Yong et al., 2025; Wang et al., 2025a). We focus on *target-language reasoning* so that users can inspect and understand intermediate steps without requiring English proficiency. To this end, we adopt a prompting strategy to encourage reasoning in the target language (Qi et al., 2025; Zhao et al., 2026a). Specifically, we insert a language-specific prefix immediately after the `<think>` token, encouraging the model to reason in the target language during both **training** and **inference**. Details are provided in §A.2, and we present language-compliance evaluation in §B.

5.3 TRAINING

Data. We use OpenThoughts (Guha et al., 2025) as our training source, which provides math reasoning problems paired with English step-by-step reference solutions. We sample 0.5K examples and translate the questions into the **17** low-resource African languages that are covered by AfriMGSM (Adelani et al., 2025).[2] The English questions and solutions are used as privileged information only for the teacher policy, while the translated questions are used in both student and teacher policies.

Implementation. Following Zhao et al. (2026b), we fix the teacher policy during training and use full-vocabulary logit distillation. We instantiate the distributional divergence with forward KL following Zhao et al. (2026b), as forward KL consistently yields the strongest gains. We independently train a separate model for each target language at each model scale, yielding 51 language-specific models across 17 languages and 3 model scales for COPSD alone. For COPSD training, we set the maximum generation length for student policy during training to 2,048 tokens. Details are provided in §G.

5.4 EVALUATION

Benchmarks. We primarily evaluate on **AfriMGSM** (Adelani et al., 2025), a human-translated version of MGSM (Shi et al., 2023) covering 17 African languages. Each language contains 250 math problems. In §6.4, we further evaluate on **PolyMath** (Wang et al., 2025c), a more challenging multilingual reasoning benchmark with problems of varying difficulty and translations reviewed and refined by language experts. For PolyMath, each language contains 125 problems per difficulty level.

Metrics. We report Pass@k (Kulal et al., 2019; Chen et al., 2021) with $k = 12$ throughout the paper. This setting probes models' reasoning potential in low-resource languages under a fixed sampling budget. For each problem, we sample 12 responses and compute whether at least one response yields the correct final answer. We instruct models to enclose their final answers in `\boxed{}`, extract the boxed content, and then compare it with the gold answer using `Math-Verify`.[3]

Checkpoint selection. To reduce checkpoint-selection bias for the main AfriMGSM results, we randomly partition the 250 problems into *50 validation problems* for checkpoint selection and *200 held-out problems* for evaluation. We select a checkpoint for each language based on validation Pass@12. To assess robustness across data splits, we repeat this selection and evaluation procedure over 10 random splits and report the mean and standard deviation of the held-out Pass@12.

Baselines. We compare COPSD against four baselines. First, we evaluate the original Qwen3 models (**Base**), which already show strong performance in high-resource languages. Second, we train Qwen3 models with **GRPO** (Shao et al., 2024) using binary outcome rewards verified against gold numerical answers. Third, we include **Trace-SFT**, where models are fine-tuned on self-generated reasoning traces produced using the teacher prompt with privileged crosslingual context. Finally, we introduce

[2] Translations are produced with `Gemini-3-Flash`. The translation prompt template is provided in §F.

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

Table 1: Pass@12 (%) on 17 low-resource AfriMGSM languages under a 4,096-token generation budget. Each language cell reports the mean (top) and standard deviation (bottom) across 10 random splits, each using 50 problems for checkpoint selection and 200 problems for evaluation. Avg. is the average of the per-language means. **Bold** indicates the highest mean within each model size and column (based on unrounded values). COPSD achieves the highest average across all model sizes.

Model	Method	AMH	EWE	HAU	IBO	KIN	LIN	LUG	ORM	SNA	SOT	SWA	TWI	VAI	WOL	XHO	YOR	ZUL	Avg.
Qwen3-1.7B	Base	15.8 ±1.2	10.2 ±0.9	12.5 ±1.2	3.6 ±0.6	6.7 ±0.7	11.1 ±1.0	5.7 ±1.2	12.3 ±1.3	7.6 ±0.4	9.1 ±0.9	14.8 ±1.2	8.7 ±1.1	5.0 ±0.6	7.4 ±0.9	6.9 ±0.5	10.0 ±1.3	9.5 ±1.3	9.2
	GRPO	15.6 ±1.2	7.4 ±0.7	13.9 ±1.0	2.7 ±0.6	9.8 ±0.8	14.9 ±1.6	7.4 ±0.9	10.6 ±1.0	9.2 ±0.8	8.2 ±1.1	12.7 ±0.9	6.5 ±0.7	2.7 ±0.3	5.6 ±0.6	7.8 ±0.8	11.8 ±1.2	10.3 ±1.4	9.2
	Trace-SFT	20.4 ±1.2	8.9 ±1.3	11.4 ±1.1	**4.1** ±0.6	8.3 ±0.7	11.4 ±1.0	9.6 ±1.1	9.3 ±1.0	11.8 ±1.3	9.5 ±0.8	15.2 ±1.3	7.8 ±0.6	6.5 ±0.7	10.6 ±1.0	6.4 ±0.8	12.3 ±0.8	11.3 ±1.3	10.3
	Translation-SFT	18.6 ±1.5	13.4 ±1.1	**18.4** ±1.5	3.1 ±0.3	14.1 ±1.1	13.8 ±0.8	13.4 ±1.2	**15.5** ±1.5	11.1 ±0.9	11.7 ±1.6	19.7 ±0.9	9.5 ±0.6	4.4 ±0.7	12.5 ±0.7	9.0 ±0.8	14.0 ±1.7	13.1 ±1.1	12.6
	COPSD	**23.4** ±1.2	**15.1** ±0.8	16.2 ±1.1	3.7 ±0.5	**14.6** ±1.2	**22.8** ±1.3	**16.5** ±1.6	13.3 ±1.3	**12.0** ±0.5	**15.1** ±1.2	**25.6** ±1.5	**13.7** ±1.1	**10.3** ±1.0	**13.9** ±1.0	**15.7** ±0.9	**14.7** ±1.3	**18.8** ±0.9	**15.6**
Qwen3-4B	Base	29.0 ±1.6	17.0 ±1.4	20.5 ±1.6	4.6 ±0.7	**17.9** ±1.2	20.6 ±0.9	14.6 ±1.8	24.5 ±1.7	14.0 ±1.2	16.8 ±1.1	47.5 ±1.5	12.5 ±1.0	**18.5** ±0.9	16.8 ±1.3	19.1 ±0.8	15.1 ±2.0	17.1 ±1.3	19.2
	GRPO	28.8 ±1.8	**19.6** ±1.5	**21.6** ±2.0	5.9 ±0.8	17.5 ±1.5	23.3 ±1.5	16.8 ±2.0	21.7 ±1.7	17.7 ±1.7	16.7 ±1.2	45.7 ±1.1	9.7 ±1.0	16.0 ±1.2	15.1 ±1.0	15.9 ±1.3	**17.5** ±1.3	17.8 ±1.9	19.2
	Trace-SFT	30.1 ±1.6	17.9 ±1.6	21.5 ±1.9	7.0 ±0.8	17.0 ±1.6	20.1 ±1.3	18.5 ±1.5	22.9 ±1.8	17.2 ±1.4	17.4 ±1.2	36.4 ±1.2	11.0 ±1.2	16.7 ±1.6	19.3 ±1.4	**19.3** ±1.6	14.1 ±1.4	16.6 ±1.5	19.0
	Translation-SFT	31.4 ±1.2	19.4 ±1.8	20.6 ±1.7	5.3 ±0.7	14.4 ±1.0	22.5 ±1.6	**19.8** ±1.8	24.3 ±1.6	**20.4** ±1.7	**20.3** ±1.5	**52.4** ±1.4	10.6 ±1.0	12.4 ±1.0	14.7 ±1.1	18.2 ±1.2	14.7 ±0.9	**20.0** ±0.9	20.1
	COPSD	**38.2** ±1.6	15.4 ±1.6	20.5 ±1.5	**8.8** ±1.0	15.7 ±1.6	**25.9** ±1.4	18.8 ±1.5	**27.1** ±1.5	17.1 ±1.6	16.4 ±1.5	47.9 ±1.6	**14.6** ±0.9	13.3 ±0.7	**20.0** ±1.0	14.2 ±1.0	17.1 ±1.0	18.4 ±1.5	**20.5**
Qwen3-8B	Base	42.3 ±1.4	9.6 ±1.0	15.7 ±1.3	3.7 ±0.5	**20.5** ±1.2	17.1 ±1.9	15.2 ±2.1	23.0 ±1.9	12.3 ±0.8	18.2 ±1.5	68.8 ±1.7	9.5 ±0.7	7.2 ±0.9	16.0 ±1.7	15.7 ±1.1	17.3 ±1.2	16.2 ±1.1	19.3
	GRPO	41.2 ±1.3	8.1 ±0.8	14.5 ±1.1	4.1 ±0.6	17.7 ±1.7	20.5 ±0.9	8.1 ±1.4	25.7 ±1.5	15.6 ±1.3	**21.8** ±1.5	**70.8** ±1.6	7.7 ±0.8	9.1 ±0.7	15.0 ±1.5	15.6 ±1.1	16.0 ±1.3	16.4 ±1.1	19.3
	Trace-SFT	43.3 ±1.4	11.4 ±1.6	18.1 ±1.6	**6.3** ±0.8	19.8 ±1.9	21.8 ±1.4	13.6 ±1.2	27.8 ±0.9	18.3 ±1.3	17.2 ±1.6	70.0 ±1.2	7.9 ±0.8	9.1 ±0.5	15.2 ±1.3	15.5 ±1.5	15.5 ±1.1	15.9 ±0.8	20.4
	Translation-SFT	39.3 ±1.0	14.4 ±1.5	18.3 ±1.5	6.2 ±0.7	19.5 ±1.8	21.4 ±1.2	15.4 ±1.6	23.4 ±1.6	20.3 ±1.2	20.1 ±1.6	70.2 ±1.7	10.3 ±1.0	**9.1** ±0.9	15.8 ±1.3	17.3 ±1.3	17.1 ±1.1	17.9 ±1.6	20.9
	COPSD	**46.3** ±1.3	**22.5** ±1.9	**21.2** ±1.2	4.1 ±0.6	18.5 ±1.8	**21.9** ±1.2	**20.1** ±1.8	**28.5** ±1.8	**20.5** ±1.5	20.0 ±0.9	66.3 ±1.7	**15.1** ±1.1	8.4 ±1.1	**18.3** ±1.3	**19.6** ±1.9	**19.6** ±1.7	**27.7** ±2.3	**23.4**

Translation-SFT, where we translate the English reference solutions into each target language and fine-tune models on these translated solutions. Further details are provided in §A.3.

5.5 Main Results

COPSD consistently improves low-resource mathematical reasoning across model scales. As shown in Table 1, COPSD achieves the highest average Pass@12 at all evaluated model sizes, improving `Qwen3-1.7B` from 9.2 to 15.6, `Qwen3-4B` from 19.2 to 20.5, and `Qwen3-8B` from 19.3 to 23.4. The gains are especially pronounced for the smallest model, with a relative improvement of approximately 69% over Base. At this scale, COPSD improves mean Pass@12 over Base in all 17 languages and achieves the highest mean among all methods in 14. This suggests that low-resource reasoning performance can be substantially improved even without target-language reasoning rationales, as long as the model is provided with dense crosslingual supervision during training. Notably, COPSD also improves performance across typologically and orthographically diverse languages, indicating that the benefit is not limited to language family or script.[4]

COPSD achieves larger average gains than outcome-based RL and offline SFT. GRPO leaves average Pass@12 largely unchanged at all three model sizes. This reflects the limited learning signal from binary outcome rewards when correct low-resource reasoning trajectories are rarely sampled. Trace-SFT yields modest average gains for `Qwen3-1.7B` and `Qwen3-8B`, but slightly underperforms Base for `Qwen3-4B`. Translation-SFT is a stronger offline baseline, achieving average scores of 12.6, 20.1, and 20.9 for the 1.7B, 4B, and 8B models, respectively. However, it requires translating full reference solutions in addition to the questions, introducing additional translation overhead. Both SFT baselines train on fixed reasoning traces using hard token targets. In contrast, COPSD conditions the teacher on privileged English information to provide token-level distributional feedback along the student's own rollouts. This supplies crosslingual supervision tailored to trajectories sampled from

[4]To assess generalization beyond the Qwen3 model family, we also conduct small-scale experiments with `DeepSeek-R1-Distill-Llama-8B` (Guo et al., 2025). We get the same pattern: COPSD improves reasoning performance across all evaluated languages; see §D for details.

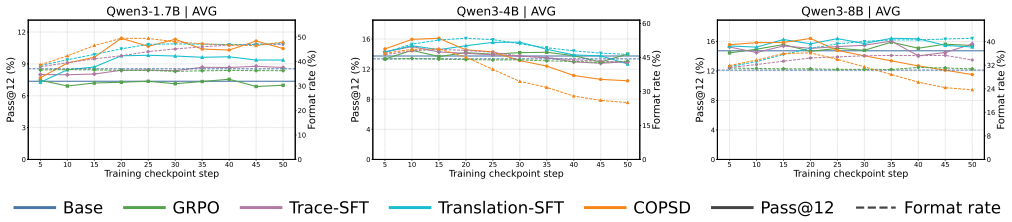
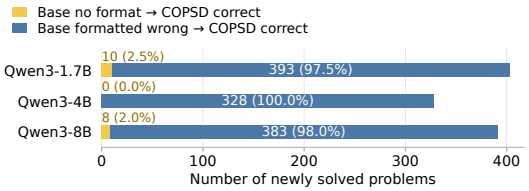


Figure 2: Average training dynamics across 17 AfriMGSM languages for `Qwen3-1.7B`, `Qwen3-4B`, and `Qwen3-8B` under a 1,024-token generation budget. Colors identify methods; solid lines show Pass@12 (left axis), and dashed lines show format rate (right axis). COPSD achieves rapid early gains, followed by relatively stable performance at 1.7B and declining Pass@12 at 4B and 8B.

Model	ρ_P^{mean}	ρ_S^{mean}	ρ_P^{pool}	ρ_S^{pool}
`Qwen3-1.7B`	0.628	0.607	0.352	0.421
`Qwen3-4B`	0.838	0.805	0.453	0.563
`Qwen3-8B`	0.728	0.693	0.232	0.479

(a) Format–accuracy correlations



(b) Newly solved problems

Figure 3: Answer-format adherence and problem-solving performance. (a) Correlations between format rate and Pass@12 across COPSD training checkpoints. ρ_P and ρ_S denote Pearson and Spearman correlations; "mean" averages per-language coefficients, while "pool" uses all language-checkpoint pairs. (b) Newly solved problem-language pairs, comparing Base with COPSD for each model. Colors distinguish whether Base produces no format-valid response or at least one format-valid response (but incorrect). Most newly solved cases are problems where Base could already produce the required format, suggesting improved reasoning beyond better format adherence.

the current student. Empirically, COPSD achieves higher average Pass@12 than both SFT baselines and GRPO at all three model sizes, supporting its effectiveness in the low-resource reasoning setting.

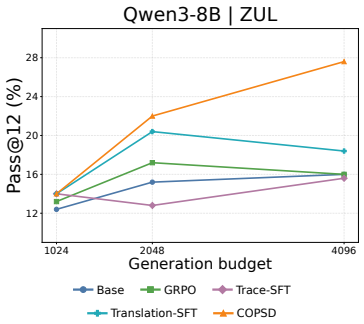
6 Further Analysis

6.1 Training Dynamics

COPSD yields rapid early gains, but later training can reduce performance. Figure 2 shows the average training dynamics across the 17 AfriMGSM languages under a 1,024-token evaluation budget.[5] Across all three model sizes, COPSD improves both **Pass@12** and **format rate** (the fraction of responses with an extractable final answer in the required `\boxed` format) during the early training stages. While `Qwen3-1.7B` eventually plateaus, `Qwen3-4B` and `Qwen3-8B` reach their best performance early and then gradually decline. This may suggest that models can quickly absorb the dense distillation signal from the teacher, but the useful signal may be limited, possibly due to weak generation capability in the low-resource languages. As a result, continued updates can begin to overfit to imperfect teacher signals. This behavior echoes prior observations that OPSD often converges rapidly (Zhao et al., 2026b). In contrast, GRPO shows no sustained upward trend in Pass@12, consistent with their limited gains in Table 1. Trace-SFT and Translation-SFT are competitive but still underperform COPSD for peak performance. This further supports our hypothesis that low-resource reasoning benefits more from dense distributional supervision.

Performance gains are associated with answer-format adherence. Figure 2 shows that Pass@12 and format rate tend to vary together during COPSD training. To further quantify this relationship, we report their correlations in Figure 3 (a). The mean per-language correlations are consistently high across model sizes, with Pearson correlations of 0.628, 0.838, and 0.728 for `Qwen3-1.7B`, `Qwen3-4B`, and `Qwen3-8B`, respectively. This suggests that low-resource reasoning failures can be partly attributed to the model's inability to produce answers in the required format within a limited token budget. However, this association does not establish that accuracy gains arise solely from

[5]We provide complete dynamics for all languages in §E.



(a) Representative language (Zulu, `Qwen3-8B`)

Model	Method	1,024	2,048	4,096
`Qwen3-1.7B`	Base	7.36	8.33 (+13.1%)	9.11 (+23.6%)
	GRPO	9.13	7.84 (-14.2%)	9.18 (+0.5%)
	Trace-SFT	10.59	9.48 (-10.4%)	10.12 (-4.4%)
	Translation-SFT	11.67	11.69 (+0.2%)	12.56 (+7.7%)
	COPSD	**13.18**	**14.47** (+9.8%)	**15.53** (+17.9%)
`Qwen3-4B`	Base	13.74	17.62 (+28.3%)	19.20 (+39.7%)
	GRPO	16.07	17.36 (+8.1%)	19.36 (+20.5%)
	Trace-SFT	16.47	17.88 (+8.6%)	19.11 (+16.0%)
	Translation-SFT	17.48	19.34 (+10.6%)	20.12 (+15.1%)
	COPSD	**18.16**	**19.51** (+7.4%)	**20.61** (+13.5%)
`Qwen3-8B`	Base	14.73	18.42 (+25.1%)	19.41 (+31.8%)
	GRPO	16.89	18.19 (+7.7%)	19.22 (+13.8%)
	Trace-SFT	17.67	18.89 (+6.9%)	20.42 (+15.6%)
	Translation-SFT	**18.61**	21.08 (+13.3%)	21.11 (+13.4%)
	COPSD	18.12	**21.18** (+16.9%)	**23.55** (+30.0%)

(b) Average across 17 languages

Figure 4: Test-time scaling of Pass@12 (%). (a) Results for Zulu (ZUL). (b) Macro-averages across 17 AfriMGSM languages; parentheses show relative changes from the corresponding 1,024-token result. **Bold** marks the highest score within each model size and budget. COPSD exhibits a clearer benefit from increased test-time computation compared to the baselines.

improved formatting. To complement this analysis, we examine newly solved problems, where the original model (Base) produces no correct response but COPSD produces at least one. As shown in Figure 3 (b), 97.5%, 100.0%, and 98.0% of these cases, respectively, already had format-valid Base responses, although all Base answers were incorrect. Thus, most gains arise on problems where formatting was already possible, suggesting improved reasoning beyond better format adherence.[6]

6.2 Ablation on Privileged Information

In COPSD, the teacher receives both the English question $x^{(H)}$ and reference solution y^*. The former helps align the low-resource input with English representation, while the latter guides correct reasoning. To assess their contributions, we compare the full setting with two ablations, providing the teacher with only $x^{(H)}$ or only y^*. We follow the setup in §5 and report results on AfriMGSM for `Qwen3-8B` with an inference-time generation budget of 4,096. in Figure 5.

Both the English problem and the reference solution are needed for the strongest supervision. Both ablations perform slightly worse than full COPSD, suggesting complementary roles for the two sources of privileged information. With only the English problem, the teacher may better interpret the low-resource input through a high-resource representation but lacks the reference solution to guide its reasoning. With only the reference solution, it has access to the correct reasoning but may struggle to align it with the low-resource problem, creating a mismatch between the student input and the teacher context. Nevertheless, both variants outperform Base, indicating that each source is useful on its own.

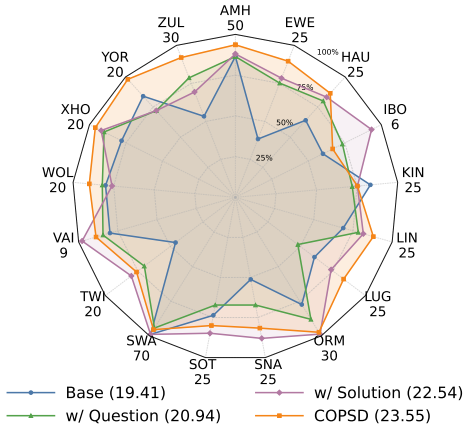


Figure 5: Ablation on privileged information. Average Pass@12 across languages is reported in parentheses in the legend. The full COPSD setting generally achieves the strongest performance across languages.

6.3 Test-Time Scaling

COPSD benefits from larger test-time generation budgets. Figure 4 shows results for one representative language, i.e., Zulu (ZUL) with `Qwen3-8B`, alongside averages across all 17 AfriMGSM languages.[7] In Figure 4 (a), COPSD starts close to the baselines at the 1,024-token budget, but separates more clearly as the budget increases. Globally, increasing the generation budget yields more consistent average improvements for the 4B and 8B models, whereas GRPO and Trace-SFT decline at 2,048 tokens for the 1.7B model. COPSD improves across all three model sizes and achieves the highest

[6] Additional analysis in §C shows that COPSD also mitigates repetitive degeneration in reasoning traces.
[7] Complete test-time scaling results for all languages and model scales are provided in §E.

average Pass@12 at both 2,048 and 4,096 tokens; Translation-SFT leads at 1,024 tokens for `Qwen3-8B`. For Qwen3-8B, COPSD improves from 18.12 to 23.55 as the generation budget increases from 1,024 to 4,096 tokens, a relative gain of 30.0%, compared with 13.8% for GRPO, 15.6% for Trace-SFT, and 13.4% for Translation-SFT. Base achieves a slightly larger relative gain of 31.8%, but reaches a lower Pass@12 of 19.41 at 4,096 tokens. These results suggest that COPSD improves not only low-resource reasoning performance, but also the model's ability to leverage longer reasoning traces in low-resource languages at inference time.

6.4 Generalization to Harder Benchmarks

To assess generalization to more challenging reasoning problems, we evaluate COPSD on **PolyMath** (Wang et al., 2025c). We select eight languages spanning different resource levels: Swahili (SWA), Telugu (TEL), Thai (THA), Russian (RUS), Bengali (BEN), Japanese (JPN), Chinese (ZHO), and Spanish (SPA). We use the same English training data source, OpenThoughts (Guha et al., 2025), as in §5. Following the same data construction procedure, we translate the English problems into each target language and use the English problems and reference solutions as privileged teacher context. For each language, we train a separate `Qwen3-4B` model with COPSD under the same training configuration as in §5. We evaluate on the low-, medium-, and high-difficulty subsets of PolyMath with a maximum generation budget of 8,192 tokens. The results are shown in Figure 6.

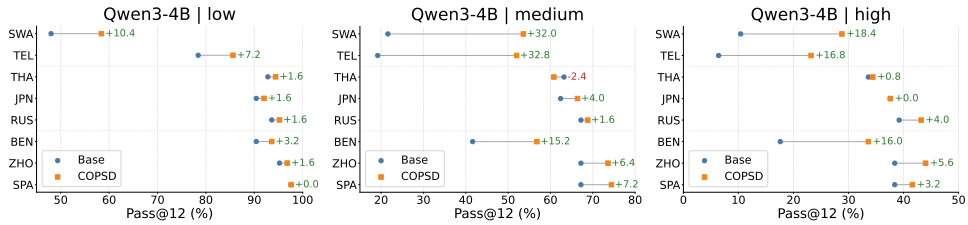


Figure 6: Pass@12 improvements of COPSD over the base model on PolyMath across low-, medium-, and high-difficulty settings. Each plot compares Base and COPSD for 8 languages of different resource levels. COPSD yields consistent gains across resource levels and difficulty levels, with especially substantial improvements for lower-resource languages, e.g., SWA and TEL.

COPSD generalizes to harder reasoning settings, with the largest gains on lower-resource languages. Across difficulty levels, COPSD improves over the base model for almost all languages, extending its benefits beyond the extremely low-resource languages in AfriMGSM. Nevertheless, we observe that the gains are particularly large for lower-resource languages. For example, on the medium-difficulty subset, COPSD improves Pass@12 by $+32.0$ points for Swahili and $+32.8$ points for Telugu, while also yielding a substantial gain of $+15.2$ points for Bengali. On the high-difficulty subset, COPSD again produces large improvements for Swahili and Telugu, with gains of $+18.4$ and $+16.8$ points, respectively. By contrast, improvements for higher-resource languages such as Japanese, Chinese, Russian, and Spanish are smaller, suggesting that these languages already benefit more from the base model's pretraining and post-training exposure, and therefore gain less from transferring English-accessible reasoning behavior. Overall, these results suggest that COPSD is most effective when the model already possesses latent reasoning ability but struggles to express that ability through lower-resource language contexts.

7 Conclusion

We introduced ***C**rosslingual **O**n-**P**olicy **S**elf-**D**istillation* (COPSD), a framework for improving multilingual mathematical reasoning, with a particular focus on low-resource languages. COPSD uses English question and reference solutions as privileged information: the student reasons from the low-resource problem alone, while the teacher, instantiated from the same model, provides dense token-level supervision on the student's own rollouts. Across 17 African languages, COPSD consistently improves over base Qwen3 models and outperforms GRPO and SFT baselines. Further analysis shows that COPSD improves format adherence, converges rapidly, mitigates repetitive degeneration, benefits from larger inference-time generation budgets, and generalizes to harder math benchmarks. Together, these results suggest that COPSD offers an effective path to mitigate the difficulty of accessing and expressing reasoning ability through underrepresented languages.

AI USE STATEMENT

The authors used ChatGPT to assist with language polishing, including grammar, clarity, and coherence, as well as minor code implementation support.[8] Gemini-3-Flash was used to translate training questions and, for Translation-SFT, English reference solutions.[9] All technical contributions, experimental design choices, and final decisions were made by the authors.

REFERENCES

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Salomey Osei, Shamsuddeen Hassan Muhammad, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. IrokoBench: A new benchmark for African languages in the age of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 2732–2757, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.139. URL https://aclanthology.org/2025.naacl-long.139/.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In Neele Falk, Sara Papi, and Mike Zhang (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 225–237, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-srw.17. URL https://aclanthology.org/2024.eacl-srw.17/.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. Long chain-of-thought reasoning across languages, 2026. URL https://arxiv.org/abs/2508.14828.

Victoria Benjamin, Emily Braca, Israel Carter, Hafsa Kanchwala, Nava Khojasteh, Charly Landow, Yi Luo, Caroline Ma, Anna Magarelli, Rachel Mirin, Avery Moyer, Kayla Simpson, Amelia Skawinski, and Thomas Heverin. Systematically analyzing prompt injection vulnerabilities in diverse llm architectures, 2024. URL https://arxiv.org/abs/2410.23308.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin

[8] https://chatgpt.com/

[9] https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/

Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13134–13156, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.729. URL https://aclanthology.org/2024.emnlp-main.729/.

Fahim Faisal, Kaiqiang Song, Song Wang, Simin Ma, Shujian Liu, Haoyun Deng, and Sathish Reddy Indurthi. Aligning multilingual reasoning with verifiable semantics from a high-resource expert model, 2025. URL https://arxiv.org/abs/2509.25543.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. A survey of multilingual reasoning in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8920–8936, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.474. URL https://aclanthology.org/2025.findings-emnlp.474/.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu,

Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12365–12394, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.826. URL https://aclanthology.org/2023.findings-emnlp.826/.

Shulin Huang, Yiran Ding, Junshu Pan, and Yue Zhang. Beyond english-centric training: How reinforcement learning improves cross-lingual reasoning in llms, 2025. URL https://arxiv.org/abs/2509.23657.

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete, and Paul Pu Liang. Learn globally, speak locally: Bridging the gaps in multilingual reasoning, 2025. URL https://arxiv.org/abs/2507.05418.

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Geunbae Lee. Why do multilingual reasoning gaps emerge in reasoning language models?, 2026. URL https://arxiv.org/abs/2510.27269.

Dayeon Ki, Kevin Duh, and Marine Carpuat. What makes good multilingual reasoning? disentangling reasoning traces with measurable features, 2026. URL https://arxiv.org/abs/2604.04720.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms?, 2026. URL https://arxiv.org/abs/2603.24472.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. Spoc: Search-based pseudocode to code. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11883–11894, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/7298332f04ac004a0ca44cc69ecf6f6b-Abstract.html.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Sufficiency/10.1214/aoms/1177729694.pdf.

Huiyuan Lai and Malvina Nissim. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12012–12026, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.649. URL https://aclanthology.org/2024.acl-long.649/.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026. URL https://arxiv.org/abs/2604.13016.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Yihong Liu, Raoyuan Zhao, Hinrich Schütze, and Michael A. Hedderich. Large reasoning models are (not yet) multilingual latent reasoners, 2026. URL https://arxiv.org/abs/2601.02996.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026.

Felix Petersen, Moritz Schubotz, Andre Greiner-Petter, and Bela Gipp. Neural machine translation for mathematical formulae. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11534–11550, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.645. URL https://aclanthology.org/2023.acl-long.645/.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle Bitterman, and Arianna Bisazza. When models reason in your language: Controlling thinking language comes at the cost of accuracy. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20279–20296, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1103. URL https://aclanthology.org/2025.findings-emnlp.1103/.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2695–2709, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.163. URL https://aclanthology.org/2023.emnlp-main.163/.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers, 2024. URL https://arxiv.org/abs/2404.04925.

Ruiyang Qin, Qingzhuo Wang, Dongrui Liu, Qiang Li, Zhihua Wei, and Wen Shen. Multilingual safety alignment via self-distillation, 2026. URL https://arxiv.org/abs/2605.02971.

Leonardo Ranaldi and Giulia Pucci. Multilingual reasoning via self-training. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 11566–11582, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.577. URL https://aclanthology.org/2025.naacl-long.577/.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. Crisp: Compressed reasoning via iterative self-policy distillation, 2026. URL https://arxiv.org/abs/2603.05433.

Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Kost, Christopher Carnahan, and Jordan Boyd-Graber. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4945–4977, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.302. URL https://aclanthology.org/2023.emnlp-main.302/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2304–2317, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.136. URL https://aclanthology.org/2024.findings-acl.136/.

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

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10015–10027, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.539. URL https://aclanthology.org/2024.acl-long.539/.

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. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Lintang Sutawika, Gokul Swamy, Zhiwei Steven Wu, and Graham Neubig. Gained in translation: Privileged pairwise judges enhance multilingual reasoning, 2026. URL https://arxiv.org/abs/2601.18722.

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

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. Reasoning transfer for an extremely low-resource and endangered language: Bridging languages through sample-efficient language understanding, 2025. URL https://arxiv.org/abs/2504.02890.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL https://aclanthology.org/2024.acl-long.845/.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schuetze. Language mixing in reasoning language models: Patterns, impact, and internal causes. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2637–2665, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.132. URL https://aclanthology.org/2025.emnlp-main.132/.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Demystifying multilingual reasoning in process reward modeling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9775–9788, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.519. URL https://aclanthology.org/2025.findings-emnlp.519/.

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. Polymath: Evaluating mathematical reasoning in multilingual contexts, 2025c. URL https://arxiv.org/abs/2504.18428.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms, 2025. URL https://arxiv.org/abs/2506.14245.

Linjuan Wu, Hao-Ran Wei, Baosong Yang, and Weiming Lu. From English to second language mastery: Enhancing LLMs with cross-lingual continued instruction tuning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23006–23023, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1121. URL https://aclanthology.org/2025.acl-long.1121/.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Language imbalance driven rewarding for multilingual self-improving. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=Kak2ZH5Itp.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation, 2026. URL https://arxiv.org/abs/2602.12125.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1028–1043, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.58. URL https://aclanthology.org/2024.acl-long.58/.

Junjie Ye, Yuming Yang, Yang Nan, Shuo Li, Qi Zhang, Tao Gui, Xuanjing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. Analyzing the effects of supervised fine-tuning on model knowledge from token and parameter levels. In Christos Christodoulopoulos, Tanmoy Chakraborty,

Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 471–513, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.25. URL https://aclanthology.org/2025.emnlp-main.25/.

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. Crosslingual reasoning through test-time scaling, 2025. URL https://arxiv.org/abs/2505.05408.

Jiaqiao Zhang, Zhoujun Li, Raoyuan Zhao, Jian Lan, Thomas Seidl, Michael A. Hedderich, Hinrich Schütze, and Yihong Liu. Beyond input understanding: Diagnosing multilingual mathematical reasoning with directed acyclic trace graphs, 2026a. URL https://arxiv.org/abs/2605.27715.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. Opsdl: On-policy self-distillation for long-context language models, 2026b. URL https://arxiv.org/abs/2604.17535.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning, 2026c. URL https://arxiv.org/abs/2510.07300.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11189–11204, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.603. URL https://aclanthology.org/2024.acl-long.603/.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. Llama beyond english: An empirical study on language capability transfer, 2024. URL https://arxiv.org/abs/2401.01055.

Raoyuan Zhao, Yihong Liu, Hinrich Schuetze, and Michael A. Hedderich. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5223–5247, Rabat, Morocco, March 2026a. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.276. URL https://aclanthology.org/2026.findings-eacl.276/.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026b. URL https://arxiv.org/abs/2601.18734.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Question translation training for better multilingual reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8411–8423, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.498. URL https://aclanthology.org/2024.findings-acl.498/.

Table 2: Language coverage of our experiments. We use the ISO 639-3 codes as language identifiers.

Code	Language	Family	Script
AMH	Amharic	Semitic	Ethiopic
EWE	Ewe	Kwa	Latin
HAU	Hausa	Chadic	Latin
IBO	Igbo	Volta-Niger	Latin
KIN	Kinyarwanda	Bantu	Latin
LIN	Lingala	Bantu	Latin
LUG	Luganda	Bantu	Latin
ORM	Oromo	Cushitic	Latin
SNA	Shona	Bantu	Latin
SOT	Sesotho	Bantu	Latin
SWA	Swahili	Bantu	Latin
TWI	Twi	Kwa	Latin
VAI	Vai	Mande	Vai
WOL	Wolof	Senegambian	Latin
XHO	Xhosa	Bantu	Latin
YOR	Yoruba	Volta-Niger	Latin
ZUL	Zulu	Bantu	Latin

Lang.	Instruction
EWE	Mesrɛ wo, bu akɔntaabu no afã afã, eye nàtsɔ wo ŋuɖoɖo mamlɛtɔ ade \boxed{} me.
HAU	Da fatan ka yi tunani mataki-mataki, kuma ka sanya amsarka ta karshe a cikin \boxed{}.
IBO	Biko tụlee ya nzọụkwụ site na nzọụkwụ, ma tinye azịza ikpeazụ gị n'ime \boxed{}.
TWI	Yɛsrɛ wo, dwene ho anammɔn biara mu, na fa wo mmuae a etwa to no hyɛ \boxed{} mu.
VAI	ꖜ ꕉꖱꕚꖶ□ ꖻꔞ ꕞꕚꖶ ꕞꕚꖶ ꕯ ꖜꖶ□ ꖻꔞ ꖜꕉꖱꕚꖶ ꕯ ꖜꖶ ꖜꖶ ꕯ ꖜꖶ ꕞꕚꖶ ꕯ ꖜꖶ ꕞꕚꖶ ꕯ ꖜꖶ ꕞꕚ

Lang.	Prefix
EWE	Le biabia me la, magɔme le susu wɔwɔ me le Eʋegbe me.
HAU	Bisa buƙata, zan fara yin tunani da Hausa.
IBO	Dị ka arịrịọ si dị, aga m amalite iche echiche n'asụsụ Igbo.
TWI	Sɛnea wɔabisa no, mɛfi ase adwene wɔ Twi mu.
VAI	Eꕇ ꕯ ꕉꕮꕾ ꕀ ꕞ ꕮꕵ ꕒꕮ ꕯ ꕤ ꕢꕢꕾ ꕀ
WOL	Ci laaj bi, dinaa tàmbali xalaat ci Wolof.
YOR	Gẹ́gẹ́ bí a ti bẹ̀rè, màá bẹ̀rẹ̀ sí í ronú ní èdè Yorùbá.
AMH	በጥያቄው መሠረት፣ በአማርኛ ማሰብ እጀምራለሁ።
KIN	Nk'uko byasabwe, ngiye gutangira gutekereza mu Kinyarwanda.
LUG	Nga bwe kisabiddwa, nja kutandika okulowooza mu Luganda.
SWA	Kwa ombi, nitaanza kufikiria kwa Kiswahili.
ORM	Gaaffii kanaan, Afaan Oromootiin yaaduu nan jalqaba.
SNA	Sekukumbirwa, ndichatanga kufunga nechiShona.
XHO	Ngokwesicelo, ndiza kuqalisa ukucinga ngesiXhosa.
ZUL	Ngokwesicelo, ngizoqala ukucabanga ngesiZulu.
SOT	Ho latela kopo, ke tla qala ho nahana ka Sesotho.
LIN	Na kolanda bosengi, nakobanda kokanisa na Lingála.

Figure 8: Prompt-hacking prefixes inserted immediately after <think> to steer the model's explicit reasoning trace in the target low-resource language.

Language-Specific Prompt Hacking Explicit language instructions alone do not always guarantee language-consistent reasoning: LLMs may still switch to English or mix languages in their reasoning traces, as observed in prior work on multilingual reasoning and language mixing (Wang et al., 2025a; Qi et al., 2025; Zhao et al., 2026a). This behavior is undesirable in our setting because it makes it difficult to compare reasoning behavior across languages and may obscure whether improvements come from better low-resource reasoning or from implicit English reasoning. To reduce such language drift, we adopt a *prompt-hacking* strategy (Schulhoff et al., 2023; Benjamin et al., 2024). Specifically, following Qi et al. (2025); Zhao et al. (2026a), we insert a target-language prefix immediately after the opening <think> tag. For example, for Swahili, we insert "*Kwa ombi, nitaanza kufikiria kwa Kiswahili.*" immediately after <think>, which means "As requested, I will begin thinking in Swahili." This prefix anchors the beginning of the reasoning trace in the target language and helps steer the model to continue reasoning in that language until the closing </think> tag. The full set of language-specific prefixes used in our experiments is listed in Figure 8.

A.3 Baseline Details

Base. We evaluate the original Qwen3 checkpoints without additional fine-tuning. At inference time, we apply the same language control trick and decoding configuration as for the trained models.

GRPO. We train models on target-language problems using a binary outcome reward. For each problem, we sample eight responses and assign a reward of 1 if the extracted final answer matches the gold answer according to `Math-Verify`, and 0 otherwise. These rewards provide the learning signal for GRPO. We allow a maximum completion length of 16,000 tokens during training. The remaining hyperparameters are listed in Table 6.

Trace-SFT. We construct supervised training data by generating complete reasoning responses using the teacher prompt from COPSD. The generation prompt includes the target-language problem, its English counterpart, and the English reference solution. Each generated response is then paired with the corresponding student prompt, which contains only the target-language problem, for supervised fine-tuning. Thus, privileged crosslingual context is used to construct the training responses, while the fine-tuned model learns to generate them from the target-language problem alone.

Translation-SFT. We translate the English reference solutions into each target language using `Gemini-3-Flash` and pair them with the corresponding target-language problems for supervised fine-tuning.[10] The translated solutions provide supervision for both the intermediate reasoning steps and the final answer. Trace-SFT and Translation-SFT use the same SFT hyperparameters, as reported in Table 7.

[10] `https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/`

B LANGUAGE COMPLIANCE

To examine whether COPSD gains arise from switching to high-resource languages such as English or French, we analyze the language of generated reasoning traces. We extract text between `<think>` and `</think>` and apply `langdetect`.[11] Since most evaluated low-resource languages are not covered by its default language profiles, we use it as a coarse detector of English/French drift. Traces classified as English or French are counted as non-compliant; other detected languages count as compliant, while failed detections count as non-compliant. This provides a permissive proxy for language compliance without requiring reliable identification of each target language. Language compliance is the percentage of generations passing this filter. Original Pass@12 measures the percentage of problems with at least one correct response among 12, whereas Compliant Pass@12 additionally requires that the correct response pass the language filter. We report macro-averages across languages.

Tables 3 and 4 report per-language compliance and aggregate performance. In general, all methods have broadly similarly good compliance rates, suggesting that our language-control prompts largely prevent systematic switching to high-resource reasoning. Taking `Qwen3-8B` as an example, all methods exhibit similar aggregate compliance rates (80.88–83.86%), with most traces passing the filter. Although COPSD has a slightly lower compliance rate than the baselines, it achieves the highest Compliant Pass@12: 20.40, compared with 16.21 for Base, 15.39 for GRPO, and 17.32 for Trace-SFT. Its performance advantage therefore persists after excluding responses with detected English/French reasoning. These results suggest that the gains cannot be explained by model switching to high-resource languages for reasoning, supporting improved problem-solving capability in low-resource languages.

Table 3: Language compliance (%) by model, method, and AfriMGSM language. Avg. is the macro-average across 17 languages. **Bold** marks the highest reported value within each model and column.

Model	Method	Avg.	AMH	EWE	HAU	IBO	KIN	LIN	LUG	ORM	SNA	SOT	SWA	TWI	VAI	WOL	XHO	YOR	ZUL
Qwen3-1.7B	Base	80.32	50.77	68.70	83.70	87.00	90.17	91.97	95.63	92.73	**90.23**	90.07	98.77	76.10	2.77	80.37	**90.13**	88.53	87.83
	GRPO	80.02	48.60	66.60	84.93	87.13	89.40	90.53	**95.83**	92.33	89.90	89.80	97.80	76.90	3.40	80.67	89.27	89.10	88.10
	Trace-SFT	**83.08**	48.67	71.70	85.27	**92.60**	**93.10**	**94.57**	93.10	93.43	89.83	**91.10**	**99.07**	85.97	**12.50**	**90.83**	88.80	**93.13**	**88.70**
	Translation-SFT	82.35	57.33	**72.60**	**91.37**	87.33	90.17	92.03	91.13	**94.50**	88.20	86.07	95.77	**87.20**	11.63	88.23	87.87	90.57	88.03
	COPSD	76.25	**75.50**	58.47	85.77	81.47	85.70	85.47	82.07	88.80	83.27	73.27	94.40	58.20	11.53	75.37	83.40	88.57	84.93
Qwen3-4B	Base	81.29	58.63	71.60	91.97	91.93	88.50	**91.33**	88.87	88.47	93.27	88.83	96.60	68.77	8.00	78.47	91.60	93.87	91.27
	GRPO	80.75	55.53	70.23	93.63	92.23	87.73	90.10	**89.07**	88.90	93.33	87.03	96.33	68.17	7.73	74.67	**92.17**	**95.20**	90.77
	Trace-SFT	**82.40**	40.03	69.97	**95.17**	**92.73**	**92.07**	90.27	88.90	89.90	**93.50**	**89.83**	**98.10**	75.13	**23.53**	**83.37**	91.80	95.03	**91.53**
	Translation-SFT	80.08	55.57	**72.33**	92.90	91.03	87.70	85.60	87.00	**90.50**	86.90	86.10	90.87	**76.60**	9.53	79.87	89.33	92.83	86.70
	COPSD	81.93	**75.27**	70.87	90.97	90.43	87.07	85.67	87.40	88.27	92.10	89.37	94.50	70.00	13.00	81.77	91.83	93.20	91.03
Qwen3-8B	Base	82.48	42.67	85.37	89.00	93.73	88.93	89.97	87.43	96.40	88.80	91.20	94.80	87.37	7.00	83.73	89.70	96.33	89.67
	GRPO	82.22	42.13	**85.90**	88.50	93.93	87.10	90.97	87.87	96.27	90.33	90.37	93.83	87.57	6.73	82.63	88.77	**96.77**	88.03
	Trace-SFT	**83.86**	31.93	83.87	89.33	**96.20**	**90.17**	**93.23**	**90.77**	96.37	**92.17**	**91.43**	94.67	**89.47**	11.70	**88.90**	**94.20**	96.57	**94.60**
	Translation-SFT	82.13	46.40	81.83	**91.93**	91.63	89.80	89.07	86.67	**97.53**	86.73	89.47	**95.00**	83.93	6.93	82.40	91.00	96.00	89.93
	COPSD	80.88	**60.50**	70.67	89.47	93.47	89.13	86.97	87.93	95.80	82.23	87.40	**95.00**	72.73	**13.97**	76.47	89.47	95.20	88.60

C REPETITION IN REASONING TRACES

Prior work has identified *repetition* as a common failure mode in multilingual reasoning, particularly in low-resource languages (Barua et al., 2026; Tran et al., 2025). Motivated by these findings, we examine whether model-generated reasoning traces exhibit repetitive degeneration and introduce a simple diagnostic metric, *repeat rate*, to quantify this behavior. Given a generated response, let $\mathcal{G}_n$ denote the multiset of all contiguous n-grams in the response, and let $\mathcal{G}_n^{\text{unique}}$ denote the set of distinct n-grams. We define the n-gram repeat rate as

$$\text{RepeatRate}_n = 1 - \frac{|\mathcal{G}_n^{\text{unique}}|}{|\mathcal{G}_n|}.$$

A higher value indicates that a larger proportion of generated n-grams are repeated. We compute this metric for $n \in \{2, 3, 4, 5, 6\}$, which allows us to capture repetition at multiple granularities, ranging from short phrase-level duplication to longer repetitive reasoning fragments.

[11] `https://github.com/Mimino666/langdetect`

Table 4: Macro-average language compliance and Pass@12 across 17 AfriMGSM languages. Compliant Pass@12 requires at least one correct response passing the language filter. **Bold** marks the highest reported value within each model.

Model	Method	Compliance rate (%)	Original Pass@12 (%)	Compliant Pass@12 (%)
Qwen3-1.7B	Base	80.32	9.11	6.87
	GRPO	80.02	9.18	7.01
	Trace-SFT	**83.08**	10.12	8.42
	Translation-SFT	82.35	12.56	10.42
	COPSD	76.25	**15.53**	**12.49**
Qwen3-4B	Base	81.29	19.20	16.19
	GRPO	80.75	19.36	16.89
	Trace-SFT	**82.40**	19.11	16.42
	Translation-SFT	80.08	20.12	17.18
	COPSD	81.93	**20.61**	**18.02**
Qwen3-8B	Base	82.48	19.41	16.21
	GRPO	82.22	19.22	15.39
	Trace-SFT	**83.86**	20.42	17.32
	Translation-SFT	82.13	21.11	18.31
	COPSD	80.88	**23.55**	**20.40**

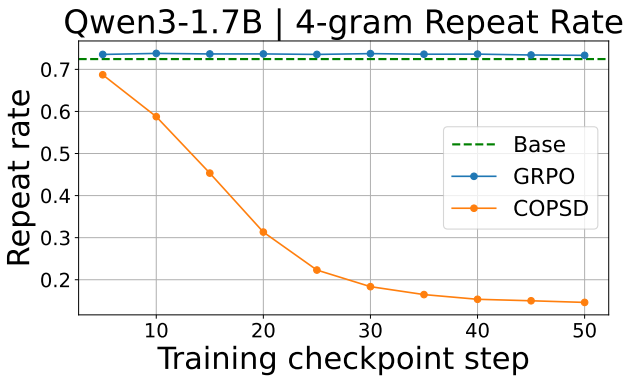


Figure 9: Average repeat rate comparison on Qwen3-1.7B with 4-grams. COPSD consistently reduces repetition compared to the base model and GRPO.

COPSD effectively mitigates repetitive degeneration in reasoning traces. Figure 9 reports the average 4-gram repeat rate of Qwen3-1.7B throughout training.[12] Compared with both the base model and GRPO, COPSD consistently maintains the lowest repeat rate across training steps. Importantly, lower repeat rates should not be interpreted simply as greater lexical diversity; rather, for reasoning in low-resource languages, they typically indicate that the model is less likely to fall into repetitive loops or produce redundant reasoning fragments. Together with the observed improvements in reasoning performance (cf. §5), this pattern suggests that COPSD encourages more coherent and structured reasoning, mitigating a failure mode in which multilingual reasoning traces collapse into meaningless or circular repetition.

D Generalization to Other Model Families

To examine generalization beyond Qwen3, we apply COPSD to DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) and evaluate Compliant Pass@12 on the medium-difficulty subset of PolyMath for Bengali (BEN), Swahili (SWA), and Telugu (TEL). As shown in Table 5, COPSD improves

[12]We report the full results for n-gram repeat rates in §E.

Table 5: Compliant Pass@12 (%) of `DeepSeek-R1-Distill-Llama-8B` on medium-difficulty PolyMath. Avg. is the macro-average across the three languages.

Method	BEN	SWA	TEL	Avg.
Base	16.0	11.2	10.4	12.5
COPSD	**19.2**	**18.4**	**13.6**	**17.1**

performance in all three languages, increasing average Compliant Pass@12 from 12.5 to 17.1, with the largest gain in Swahili. These results provide evidence that the benefits of COPSD extend beyond Qwen3 to a Llama-based reasoning model.

E Complete Results

We provide the complete per-language training dynamics for all three model sizes in Figure 11, Figure 12, and Figure 13. These figures complement the averaged results in Figure 2 and show that the main trends are broadly consistent across languages: COPSD typically improves Pass@12 and format rate within the early training steps, while GRPO often exhibits flatter or more unstable trajectories. At the same time, the language-level plots reveal substantial variation across languages, suggesting that the effectiveness and saturation point of COPSD depend on both model scale and target-language generation quality.

Figure 14, Figure 15, and Figure 16 report the complete per-language test-time scaling results under generation budgets of 1024, 2048, and 4096 tokens for all three model sizes. Overall, COPSD tends to outperform the base and GRPO-trained models across budgets, although the magnitude of improvement varies by language and model size. The benefits of increased generation budget are more consistent for larger models, especially `Qwen3-8B`, supporting the observation in §6.3 that effective crosslingual test-time scaling requires sufficient model capacity.

Figure 10 presents repeat rate comparisons for $n = 2$ to 6. Across all settings, COPSD consistently achieves lower repeat rates than both the base model and GRPO. This pattern holds across different model scales and n-gram granularities, indicating that the reduction in repetition is robust and COPSD effectively improves the quality of low-resource reasoning.

F Prompt Template

This section summarizes the three prompt templates used in our experiments: (i) the translation prompt for constructing low-resource training questions from OpenThoughts (Guha et al., 2025), (ii) the student-policy prompt, and (iii) the teacher-policy prompt. The student-policy and the teacher-policy are adapted from prompts used by Zhao et al. (2026b). For the student and teacher policies, we illustrate the instantiated prompts using Swahili as an example. For readability, we additionally provide English reference translations of the Swahili prompts. The complete prompt templates for all languages are available in our GitHub repository.[13]

F.1 Translation Prompt

We use a translation prompt to convert English mathematical problems into target languages from 17 low-resource African languages. This prompt is used with `Gemini-3-Flash` to translate only the problem text while preserving mathematical content, numbers, and LaTeX expressions.[14]

F.2 Student-Policy Prompt

The student policy receives only the low-resource problem and a language-specific instruction asking the model to reason step by step in the target language. Figure 18 shows the instantiated prompt for Swahili, together with an English reference translation.

[13] Hidden for anonymity.

[14] `https://aistudio.google.com/models/gemini-3`

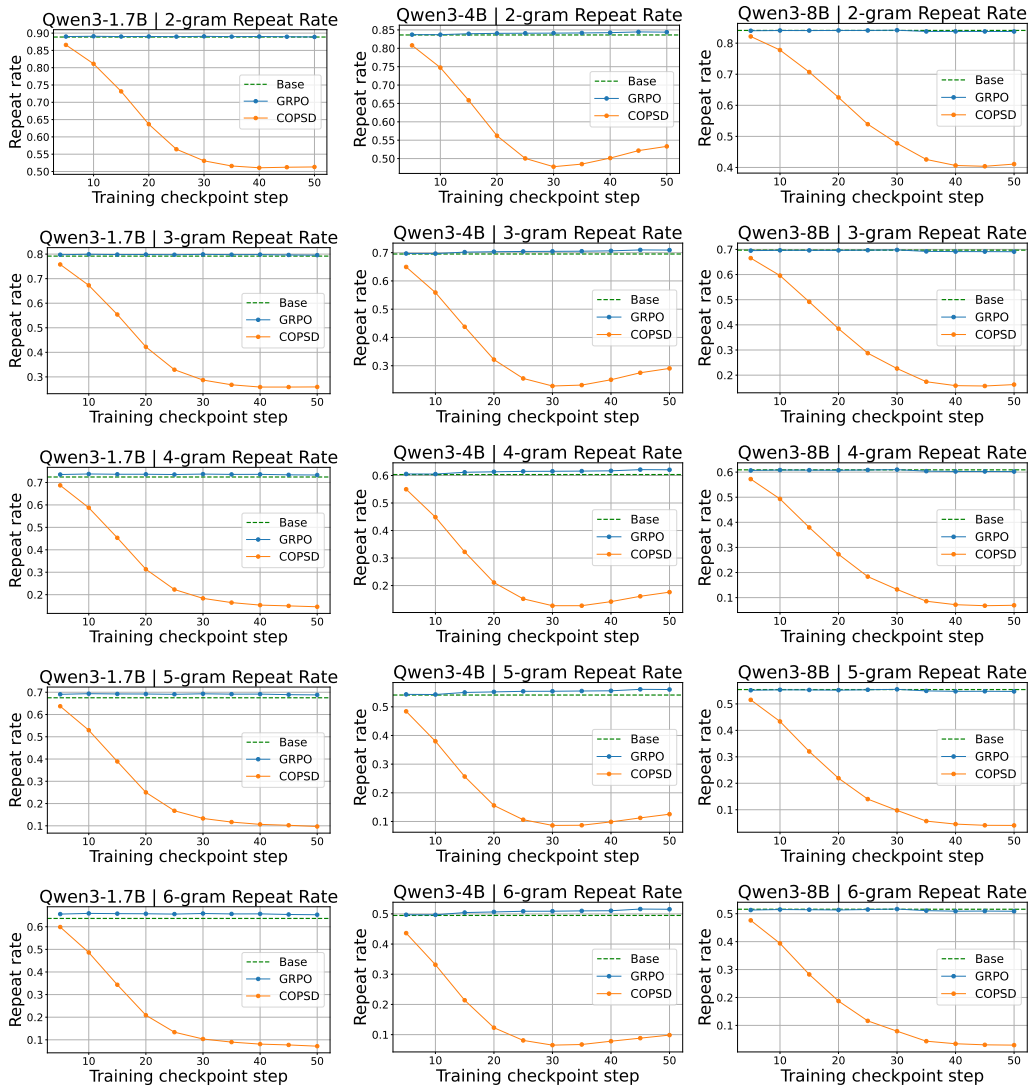


Figure 10: Repeat rate across different n-gram settings ($n = 2$ to 6) and model sizes. COPSD consistently reduces repetition compared to both the base and GRPO.

F.3 Teacher-Policy Prompt

The teacher policy is given privileged crosslingual information, including the target-language problem, the English translation of the problem, and the English reference solution. It is then asked to solve the original low-resource problem in the target language. Figure 19 shows the teacher prompt instantiated for Swahili, together with an English reference translation.

For all other languages, the same prompt structure is used with language-specific instructions, labels, and reasoning prefixes. The complete prompt templates for every language in our experiments are provided in our GitHub repository.

G Environment and Hyperparameters

We largely follow the training configuration of Zhao et al. (2026b) for both GRPO and OPSD-style training. The main difference is that we set the maximum completion length for COPSD to 2048 tokens, instead of the 1024-token budget used in the original OPSD setup. Unlike Zhao et al. (2026b), we enable thinking mode for both the student and teacher policies. This is necessary for eliciting

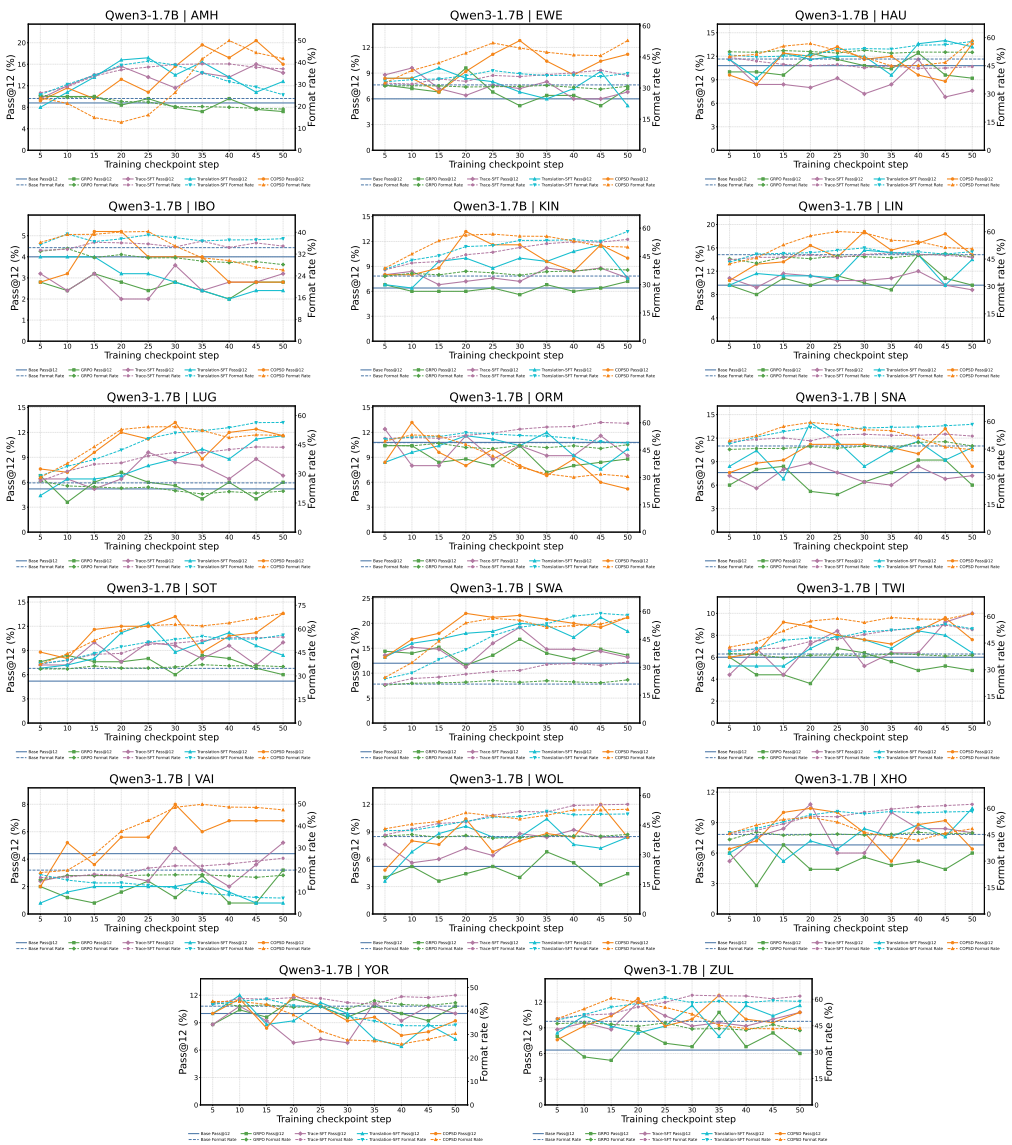


Figure 11: Per-language training dynamics for `Qwen3-1.7B` across all African languages under a 1024-token generation budget. Solid lines show Pass@12 and dashed lines show format rate.

language-specific reasoning traces, as our language-control strategy inserts a target-language prefix immediately after the `<think>` token, as described in §A.2.

We train a separate model for each language, model scale, and training method. For the main COPSD experiments, this results in $17 \times 3 = 51$ models; we use the same language–scale grid for GRPO and SFT baselines. All experiments are conducted on either 8 NVIDIA A100 GPUs or 4 NVIDIA H200 GPUs. We use LoRA (Hu et al., 2022) for parameter-efficient fine-tuning, AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) as the optimizer, and bfloat16 precision for all training runs. By default, COPSD uses full-vocabulary logit distillation with a fixed teacher policy. For COPSD, GRPO, and SFT, we save checkpoints every 5 training steps. Table 6 summarizes GRPO and COPSD hyperparameters, while Table 7 reports the SFT (Trace-SFT and Translation-SFT) configuration.

For evaluation, we use the same decoding configuration for all models and methods to ensure fair comparison, as shown in Table 8. We enable thinking mode and sample 12 responses per problem with temperature 1.0 and top-$p = 0.95$. For AfriMGSM, we evaluate under maximum new-token budgets of 1,024, 2,048, and 4,096 tokens. These budgets are sufficient for AfriMGSM because

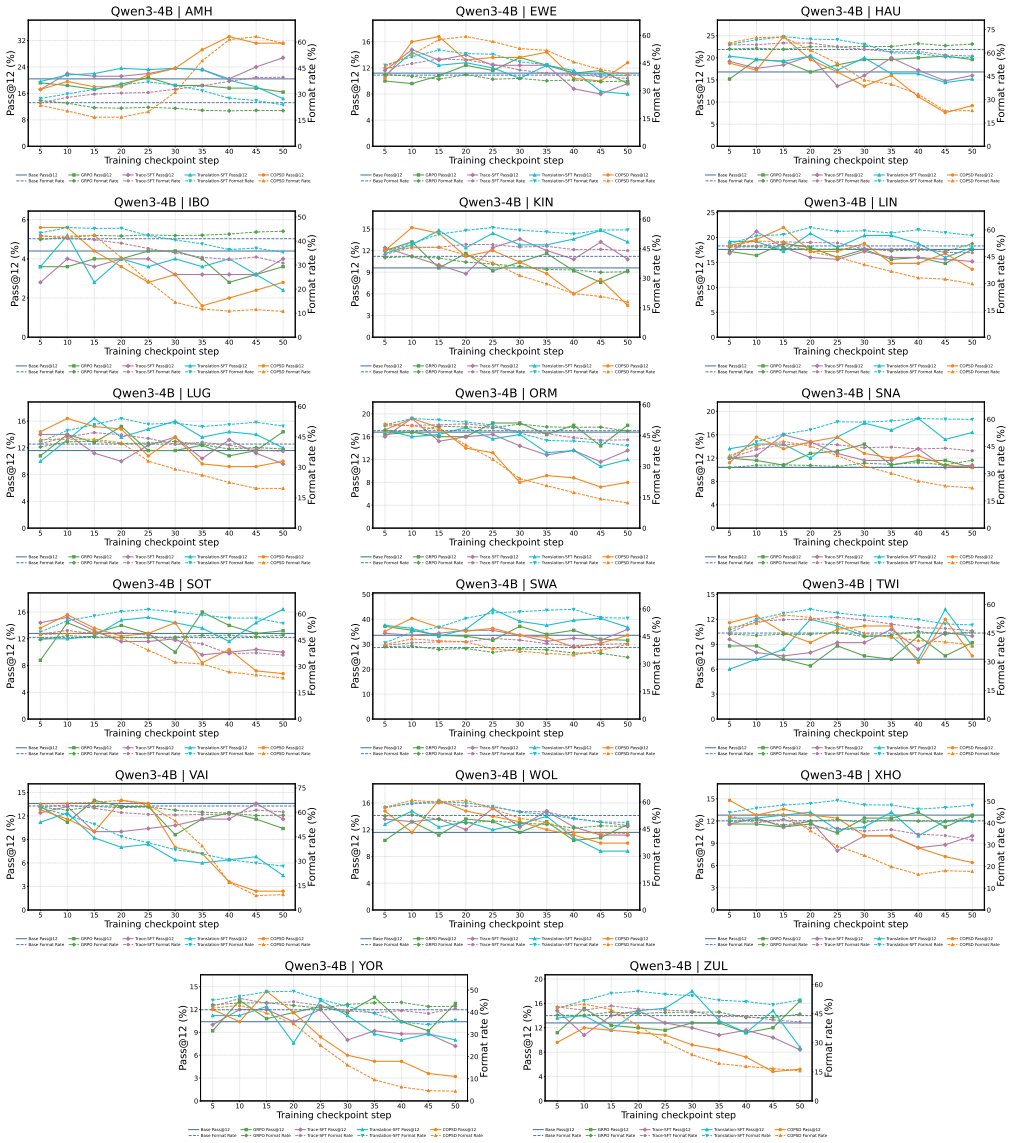


Figure 12: Per-language training dynamics for `Qwen3-4B` across all African languages under a 1024-token generation budget. Solid lines show Pass@12 and dashed lines show format rate.

the benchmark consists of relatively short mathematical reasoning problems. Final answers are extracted from `\boxed{}` and verified as described in §5. For checkpoint selection on AfriMGSM, we randomly partition the 250 problems in each language into 50 validation problems and 200 held-out evaluation problems. For each trained method and model size, we select a checkpoint separately for each language based on validation Pass@12 and evaluate it on the corresponding held-out problems. We repeat this procedure over 10 random splits and report the mean and standard deviation of held-out Pass@12. This protocol applies to the main AfriMGSM experiments (§5). For subsequent analyses, we select the checkpoint with the highest Pass@12 on the full set of 250 problems for each language, trained method, and model size, and report its performance on the same set.

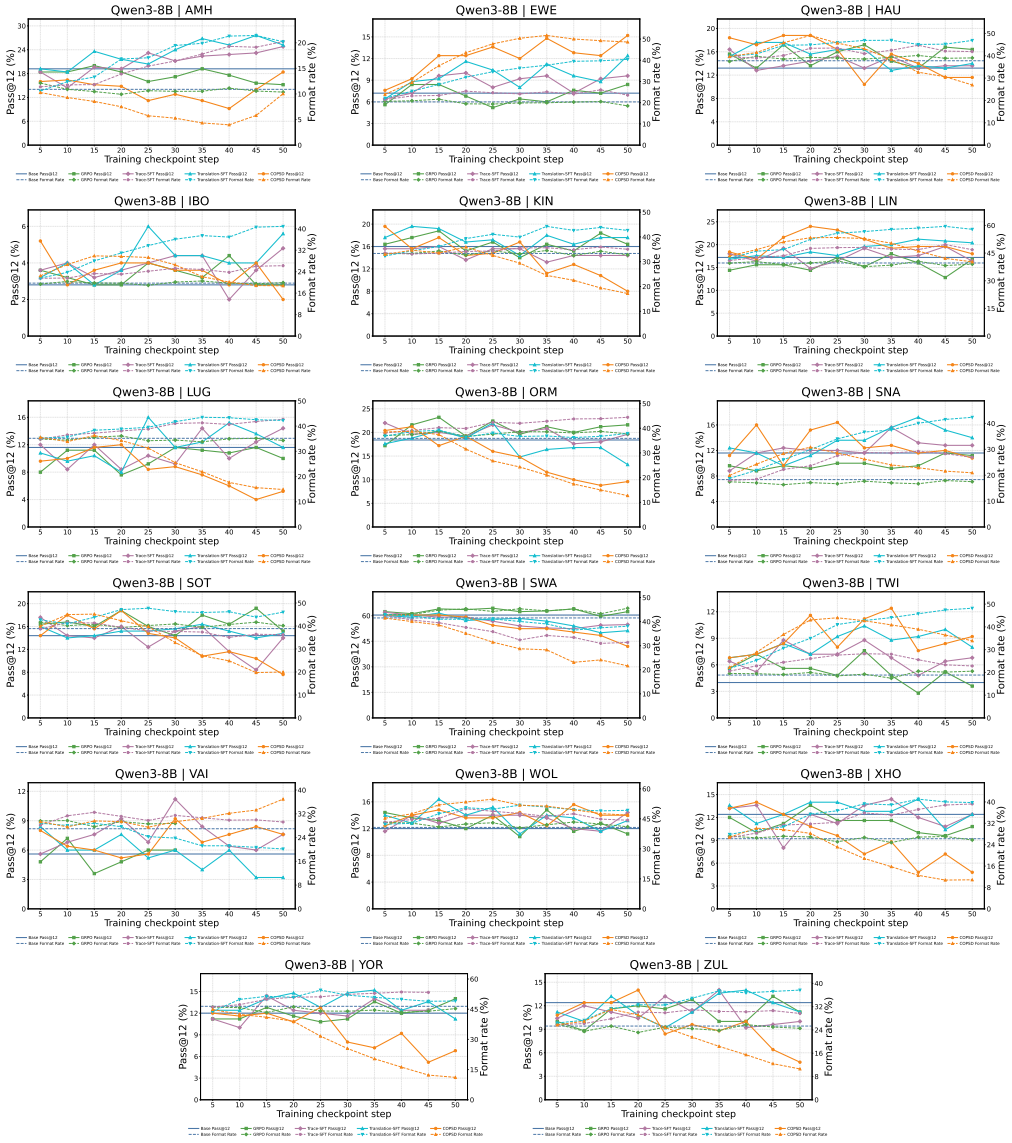


Figure 13: Per-language training dynamics for `Qwen3-8B` across all African languages under a 1024-token generation budget. Solid lines show Pass@12 and dashed lines show format rate.

Table 6: Training hyperparameters for GRPO and COPSD. We follow the original OPSD configuration (Zhao et al., 2026b) except that COPSD uses a maximum completion length of 2048 tokens to allow more supervision on low-resource generations.

Parameter	**GRPO**	**COPSD**
Learning Rate	5×10^{-6}	5×10^{-6}
Effective Batch Size	32	32
LoRA Rank (r)	64	64
LoRA Alpha (α)	128	128
Max Completion Length	16,000	2,048
Generations per Prompt	8	1
Sampling Temperature	1.2	1.1
KL Coefficient (β)	0.0	–
Training Steps	100	100

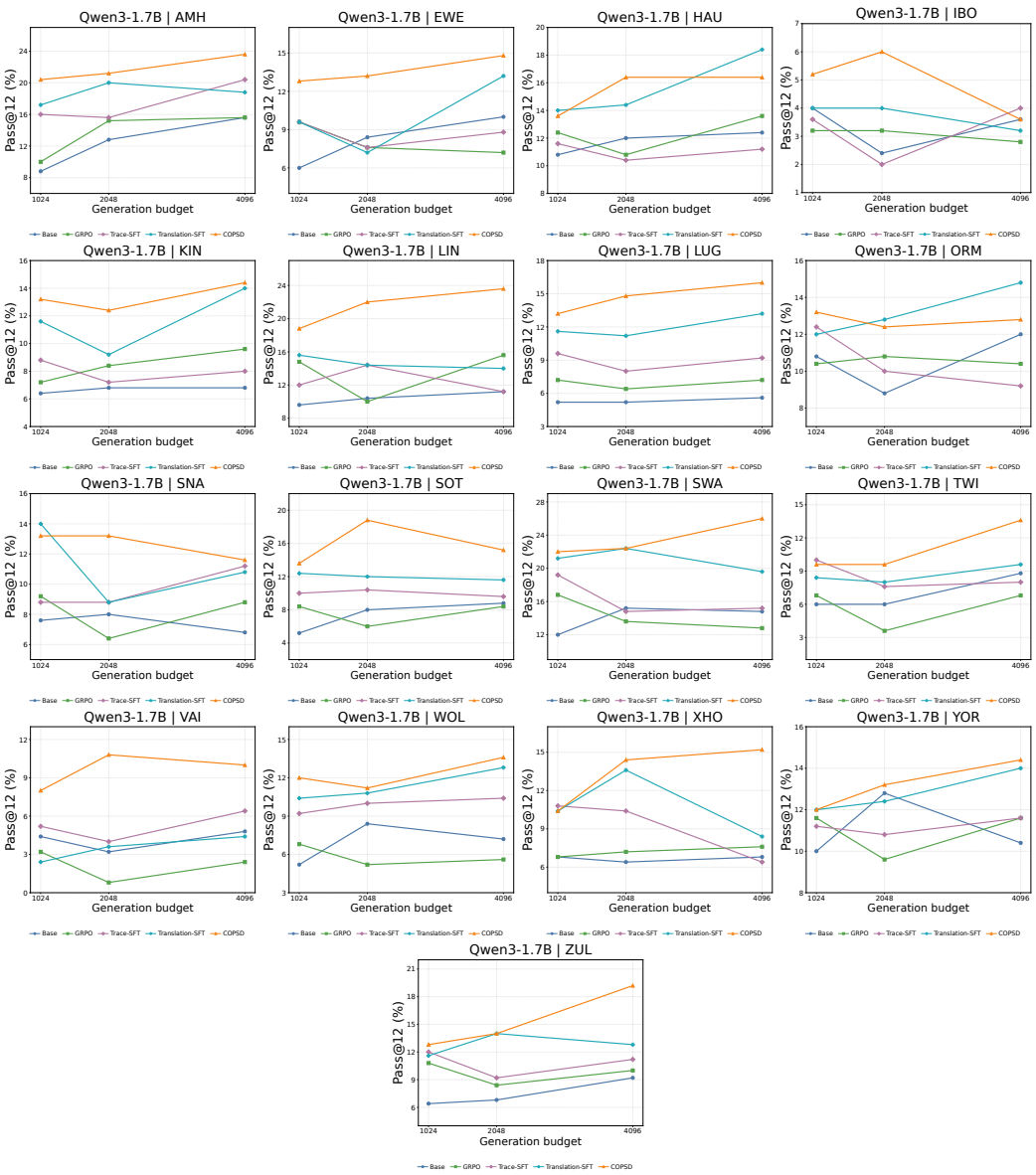


Figure 14: Per-language test-time scaling results on Pass@12 for `Qwen3-1.7B` across all African languages, under generation budgets of 1024, 2048, and 4096 tokens. Overall, the trends are mixed across languages, but COPSD generally achieves stronger performance than both the Base model and GRPO under different generation budgets.

Table 7: Training hyperparameters shared by Trace-SFT and Translation-SFT. Each training sequence concatenates the student prompt with either a self-generated reasoning response or a translated reference solution, with a maximum sequence length of 8,192 tokens.

Parameter	**SFT**
Learning Rate	5×10^{-6}
Effective Batch Size	32
LoRA Rank (r)	64
LoRA Alpha (α)	128
Max Sequence Length	8,192
Training Steps	100

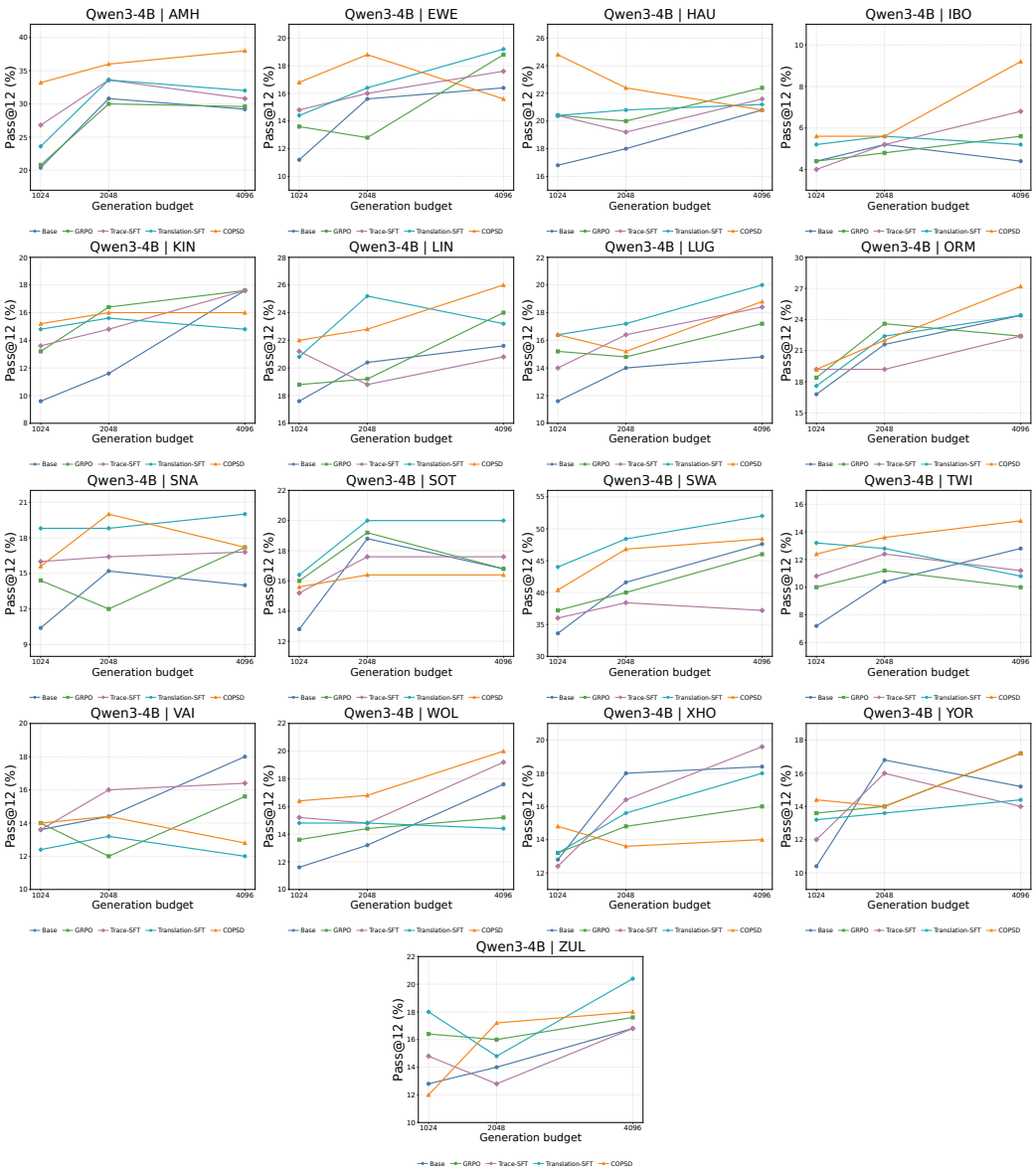


Figure 15: Per-language test-time scaling results on Pass@12 for `Qwen3-4B` across all African languages, under generation budgets of 1024, 2048, and 4096 tokens. Overall, the trends are mixed across languages, but COPSD generally achieves stronger performance than both the Base model and GRPO under different generation budgets.

Table 8: Inference hyperparameters used for evaluation. We use the same decoding configuration for all models and methods.

Parameter	**Value**
Max New Tokens	1,024, 2,048, or 4,096
Thinking Mode	Enabled
Top-p	0.95
Top-k	-1
Min-p	0.0
Presence Penalty	0.0
Samples per Prompt	12
Temperature	1.0

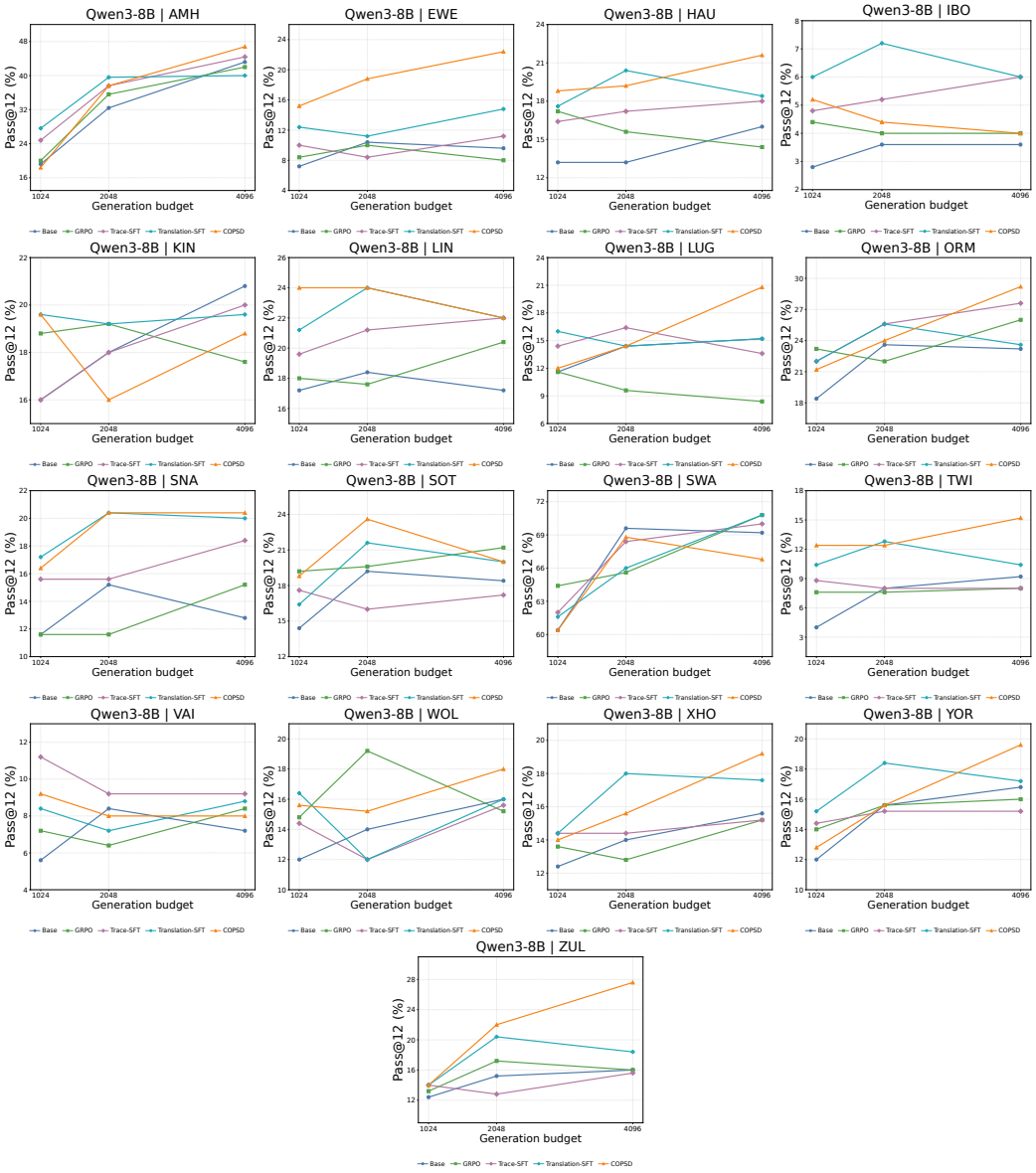


Figure 16: Per-language test-time scaling results on Pass@12 for `Qwen3-8B` across all African languages, under generation budgets of 1024, 2048, and 4096 tokens. Overall, the trends are mixed across languages, but COPSD generally achieves stronger performance than both the Base model and GRPO under different generation budgets.

Translation prompt template (used with Gemini-3-Flash)

```
Translate the following competition math problem from English into {language_name}.
Requirements:
- Return only the translated problem text.
- Do not add explanations, notes, quotation marks, or formatting wrappers.
- Preserve all numbers exactly.
- Preserve all LaTeX expressions exactly as they appear.
- Preserve the meaning and the final asked quantity exactly.
- Do not solve the problem.
English problem:
{problem}
```

Figure 17: Translation prompt template used to translate English training questions into a target language. The prompt explicitly constrains the model to preserve mathematical content and return only the translated problem text.

Student-policy prompt (Swahili)

Swali: *[problem_target]*

Tafadhali fikiri hatua kwa hatua, na uweke jibu lako la mwisho ndani ya `\boxed{}`.

English reference translation

Question: *[problem_target]*

Please think step by step, and place your final answer inside `\boxed{}`.

Figure 18: Student-policy prompt. The left panel shows the instantiated Swahili prompt used in training and inference; the right panel provides an English reference translation.

Teacher-policy prompt (Swahili)

Swali: *[problem_target]*

Tafsiri ya Kiingereza ya swali: *[problem_english]*

Suluhisho sahihi la rejeleo kwa Kiingereza:
`=== Mwanzo wa Suluhisho la Rejeleo ===`
[solution_english]
`=== Mwisho wa Suluhisho la Rejeleo ===`

Baada ya kusoma suluhisho la rejeleo la Kiingereza hapo juu, hakikisha umeelewa kweli mantiki ya kila hatua—usilinakili wala kulifafanua upya tu. Sasa, kwa kutumia maneno yako mwenyewe na hoja huru, tatua swali la asili kwa Kiswahili. Fikiri hatua kwa hatua, jaribu mbinu tofauti, na usiogope kurudi nyuma au kufikiria upya ikiwa jambo fulani halifanyi kazi:

Tafadhali fikiri hatua kwa hatua kwa Kiswahili, na uweke jibu lako la mwisho ndani ya `\boxed{}`.

English reference translation

Question: *[problem_target]*

English translation of the question: *[problem_english]*

Correct reference solution in English:
`=== Begin Reference Solution ===`
[solution_english]
`=== End Reference Solution ===`

After reading the English reference solution above, make sure you truly understand the logic of each step—do not simply copy or paraphrase it. Now, using your own words and independent reasoning, solve the original question in Swahili. Think step by step, try different approaches, and do not be afraid to backtrack or rethink if something does not work:

Please think step by step in Swahili, and place your final answer inside `\boxed{}`.

Figure 19: Teacher-policy prompt. The teacher receives privileged information, including the English translation of the problem and the English reference solution, before solving the original low-resource problem. The left panel shows the instantiated Swahili prompt; the right panel provides an English reference translation.