Crosslingual On-Policy Self-Distillation for Multilingual Reasoning

Anonymous ACL submission

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 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, strengthens test-time scaling, and generalizes to harder multilingual reasoning benchmarks, with especially large gains for lower-resource languages.

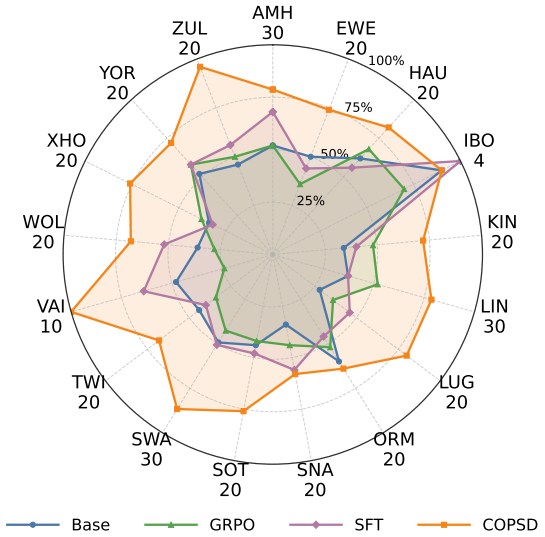

Figure 1: Radar comparison of `Qwen3-1.7B` performance on AfriMGSM under a 4096-token generation budget. Each axis corresponds to one of the 17 low-resource African languages, with axis-specific scaling based on the maximum observed performance for that language. COPSD consistently outperforms other baselines, including GRPO and SFT, across 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 pretraining and are rarely represented in high-quality reasoning supervision during post-training (Qin et al., 2024; Yang et al., 2025b). As a result, a model may possess the latent ability to solve a problem, yet fail to access that ability 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 target languages and then performing supervised fine-tuning (SFT) (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 distribu-

tion 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 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.

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., 2026a; 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 (see Figure 1). 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 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 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 and 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., 2026a; Kim et al., 2026; Sang et al., 2026). Our work extends OPSD to the multilingual setting, enabling the model to transfer its English-accessible reasoning behavior to low-resource languages and offering an effective, novel approach to improving low-resource reasoning.

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), 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 and Nissim, 2024), *self-training* (Ranaldi and Pucci, 2025; Sutawika et al., 2026), and *reinforcement learning* (She et al., 2024; Ranaldi and Pucci, 2025; Wang et al., 2025b; Huang et al., 2025; Faisal et al., 2025; Zhang et al., 2026b). However, these approaches typically require translated reasoning rationales or sparse outcome rewards. In contrast, 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., 2026a). Instead of distilling knowledge from an external model (Agarwal et al., 2024; Lu and Lab, 2025), OPSD instantiates the same model as both a student and a teacher under different conditioning contexts. Given a reasoning dataset $\mathcal{D} = \{(x, y^*)\}$, where x is a problem and y^* is privileged information such as a reference solution, OPSD defines two policies from the same model p_θ:

$$p_S(\cdot \mid x) \triangleq p_\theta(\cdot \mid x),$$
$$p_T(\cdot \mid x, y^*) \triangleq p_\theta(\cdot \mid x, y^*).$$

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 $\hat{y}_{<n}$:

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

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 and Leibler, 1951). The overall OPSD objective is

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

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

3.4 Discussion

OPSD is attractive because: (i) It learns from on-policy student-generated trajectories, exploits privileged information, and avoids the need for an external teacher. (ii) Compared with SFT/off-policy distillation, it reduces train-test mismatch by training on the student's own generations. (iii) Compared with outcome-based RL, it provides dense teacher feedback over intermediate reasoning steps rather than relying only on sparse final-answer rewards.

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*. During training, the student must reason from the low-resource problem alone, while the teacher is given additional high-resource, English context that helps elicit a stronger reasoning distribution from the same model, as shown in Figure 2. This allows the model to transfer its own English-accessible reasoning behavior to low-resource languages without relying on an external teacher or target-language rationales.

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 high-resource language

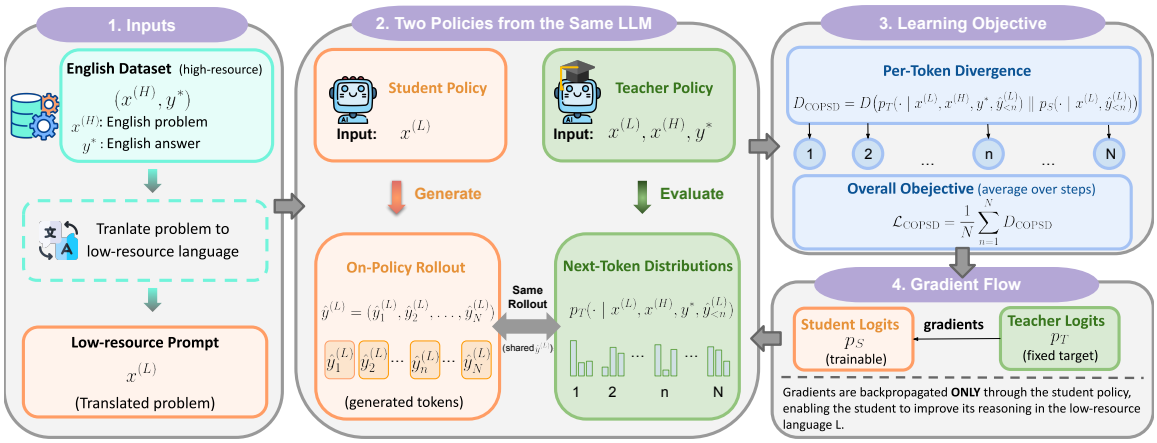

Figure 2: Overview of COPSD. Each problem is translated into a low-resource language as the student's input. The same LLM acts as both student and teacher: the student generates an on-policy rollout, while the teacher evaluates it with privileged English context and the reference solution. By minimizing per-token divergence along the rollout, COPSD transfers English-accessible reasoning behavior to improve reasoning in low-resource languages.

counterpart, and y^* is the reference solution in the high-resource language. 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).

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 receives privileged crosslingual information:

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

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 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:

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

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

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

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

5 Experiments

5.1 Models

We conduct experiments with Qwen3 family models (Yang et al., 2025a) of three sizes: `Qwen3-1.7B`, `Qwen3-4B`, and `Qwen3-8B`. Qwen3 models are pre-trained on multilingual corpora and further post-trained with SFT and RL 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). Since our goal is to improve reasoning in specific low-resource languages, we control the reasoning language with a prompt-design strategy (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**. Further details are provided in §A.2.

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

[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.

Model	Method	AMH	EWE	HAU	IBO	KIN	LIN	LUG	ORM	SNA	SOT	SWA	TWI	VAI	WOL	XHO	YOR	ZUL	Avg.
Qwen3-1.7B	Base	15.60	10.00	12.40	3.60	6.80	11.20	5.60	12.00	6.80	8.80	14.80	8.80	4.80	7.20	6.80	10.40	9.20	9.11
	GRPO	15.60	7.20	13.60	2.80	9.60	15.60	7.20	10.40	8.80	8.40	12.80	6.80	2.40	5.60	7.60	11.60	10.00	9.18
	SFT	20.40	8.80	11.20	**4.00**	8.00	11.20	9.20	9.20	11.20	9.60	15.20	8.00	6.40	10.40	6.40	11.60	11.20	10.12
	COPSD	**23.60**	**14.80**	**16.40**	3.60	**14.40**	**23.60**	**16.00**	**12.80**	**11.60**	**15.20**	**26.00**	**13.60**	**10.00**	**13.60**	**15.20**	**14.40**	**19.20**	**15.53**
Qwen3-4B	Base	29.20	16.40	20.80	4.40	**17.60**	21.60	14.80	24.40	14.00	16.80	47.60	12.80	**18.00**	17.60	18.40	15.20	16.80	19.20
	GRPO	29.60	**18.80**	**22.40**	5.60	**17.60**	24.00	17.20	22.40	**17.20**	16.80	46.00	10.00	15.60	15.20	16.00	**17.20**	17.60	19.36
	SFT	30.80	17.60	21.60	6.80	**17.60**	20.80	18.40	22.40	16.80	**17.60**	37.20	11.20	16.40	19.20	**19.60**	14.00	16.80	19.11
	COPSD	**38.00**	15.60	20.80	**9.20**	16.00	**26.00**	**18.80**	**27.20**	**17.20**	16.40	**48.40**	**14.80**	12.80	**20.00**	14.00	**17.20**	**18.00**	**20.61**
Qwen3-8B	Base	43.20	9.60	16.00	3.60	**20.80**	17.20	15.20	23.20	12.80	18.40	69.20	9.20	7.20	16.00	15.60	16.80	16.00	19.41
	GRPO	42.00	8.00	14.40	4.00	17.60	20.40	8.40	26.00	15.20	**21.20**	**70.80**	8.00	8.40	15.20	15.20	16.00	16.00	19.22
	SFT	44.40	11.20	18.00	**6.00**	20.00	**22.00**	13.60	27.60	18.40	17.20	70.00	8.00	**9.20**	15.60	15.20	15.20	15.60	20.42
	COPSD	**46.80**	**22.40**	**21.60**	4.00	18.80	**22.00**	**20.80**	**29.20**	**20.40**	20.00	66.80	**15.20**	8.00	**18.00**	**19.20**	**19.60**	**27.60**	**23.55**

Table 1: Pass@12 performance on 17 low-resource AfriMGSM languages under a 4,096-token generation budget. **Bold** values indicate the best result among Base, GRPO, SFT, and COPSD for each model size and language. COPSD outperforms baselines on most languages, with large gains for `Qwen3-1.7B` and `Qwen3-8B`.

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 for the teacher policy, while the translated questions are used for the student policy.

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 reverse KL. For all models, we set the student maximum generation length to 2048 tokens. Details are provided in §E.

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 reasoning problems. In §6.4, we further evaluate on **PolyMath** (Wang et al., 2025c), a more challenging multilingual reasoning benchmark with problems of varying difficulty. For PolyMath, each language contains 125 questions.

Metrics. We report Pass@k (Kulal et al., 2019; Chen et al., 2021) with $k = 12$ throughout the paper. 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]

Baselines. We compare COPSD against three 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 an **SFT** baseline, where models are fine-tuned on self-generated reasoning traces produced from the teacher prompt with privileged crosslingual context.

5.5 Results and Discussion

COPSD consistently improves low-resource mathematical reasoning across model scales. As shown in Table 1, COPSD achieves the best average Pass@12 performance for all evaluated model sizes, improving `Qwen3-1.7B` from 9.11 to 15.53, `Qwen3-4B` from 19.20 to 20.61, and `Qwen3-8B` from 19.41 to 23.55. The gains are especially pronounced for the smaller model, where COPSD improves performance on nearly every language and yields a relative improvement of over 70% in average Pass@12 over the base model. 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.

Outcome-based RL and offline SFT provide limited gains in low-resource languages, while COPSD offers a denser and more reliable learning signal. GRPO yields only modest improvements for `Qwen3-1.7B` and `Qwen3-4B`, and even underperforms the base model for `Qwen3-8B`. This suggests that binary outcome rewards provide weak supervision when correct low-resource reasoning trajectories are rarely sampled. SFT on self-generated teacher traces can transfer some privileged-context behavior, but it uses discrete sampled traces as hard targets, making the supervision less flexible and less informative than the full teacher distribution. In contrast, COPSD provides token-level distributional feedback along the stu-

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

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

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

Table 2: Correlation between format rate and Pass@12 for COPSD during training. ρ_P and ρ_S denote Pearson and Spearman correlations, respectively. The mean correlation averages coefficients computed independently for each language trajectory, while the pooled correlation is computed over all language-step pairs.

dent's own rollouts. By conditioning the teacher on privileged English information, the same model can serve as an effective crosslingual teacher, giving dense on-policy guidance for improving reasoning in the target low-resource language.

6 Further Analysis

6.1 Training Dynamics

COPSD improves performance rapidly in early steps Figure 3 shows the average training dynamics across the 17 languages under the 1,024-token evaluation budget.[4] Across all model sizes, COPSD improves both Pass@12 and format rate in the early training steps.[5] While Qwen3-1.7B eventually plateaus, Qwen3-4B and Qwen3-8B reach their best performance early and then gradually decline. This suggests 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 begin to overfit to imperfect teacher signals. This behavior echoes prior observations that OPSD often converges rapidly (Zhao et al., 2026b). In contrast, GRPO and SFT show no clear upward trend in Pass@12, consistent with their limited gains in Table 1. This further supports our hypothesis that low-resource reasoning benefits more from dense distributional supervision than from either sparse binary rewards or hard-token targets derived from sampled traces.

Performance gains are closely tied to answer-format adherence. Figure 3 suggests a strong association between Pass@12 and format rate. To further quantify this relationship, we report their correlations in Table 2. The mean per-language correlations are consistently high across model sizes, with Pearson correlations of 0.628, 0.838, and 0.728

[4]We provide complete dynamics for all languages in §C.

[5]Format rate is the fraction of responses that provide a valid final answer in the required \boxed{} format.

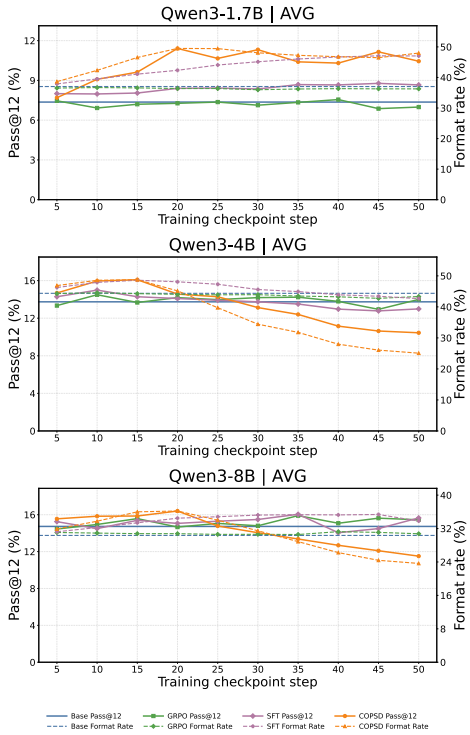

Figure 3: Average training dynamics across languages for Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. Solid lines show Pass@12 and dashed lines show format rate. Overall, COPSD converges quickly and often reaches its best performance within only a few training steps, while GRPO shows no clear improvement trend over training.

for Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. Although the pooled correlations are lower, they remain positive, indicating that the relationship holds both within individual language learning trajectories and across all language–checkpoint pairs. This suggests that low-resource reasoning failures can be partly caused by the model's inability to produce answers in the required format within a limited token budget. The decline in format rate for larger models (4B and 8B) after early COPSD checkpoints therefore helps explain the corresponding drop in Pass@12 in Figure 3. These observations motivate our next analysis on test-time scaling (§6.3), where we examine whether larger generation budgets can recover or amplify the reasoning gains learned through COPSD.

6.2 Ablation on Privileged Information

In COPSD, the teacher observes both the English translation of the problem $x^{(H)}$ and the English reference solution y^*. This design provides two complementary signals: the English problem helps the teacher semantically align the low-resource input with a high-resource representation, while the reference solution guides the teacher toward correct

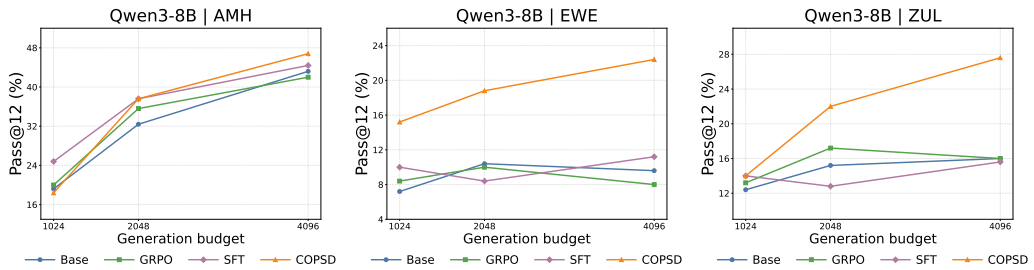

Figure 4: Test-time scaling results for three representative languages: Amharic (AMH), Ewe (EWE), and Zulu (ZUL). COPSD exhibits a clearer benefit from increased test-time computation compared to the baselines.

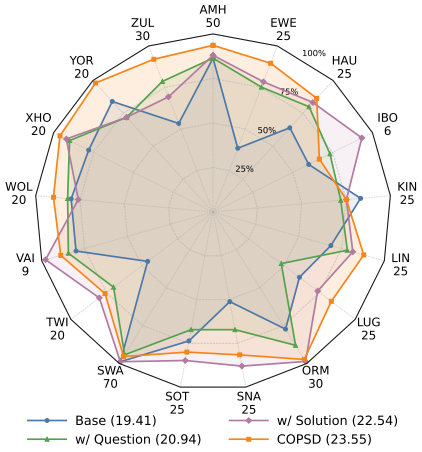

Figure 5: Ablation on privileged information. Average Pass@12 across languages is reported in parentheses in the legend. The full COPSD setting generally achieves the strongest performance, indicating that the English problem and the English reference solution provide complementary supervision for low-resource reasoning.

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%)
	SFT	10.59	9.48 (-10.4%)	10.12 (-4.4%)
	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%)
	SFT	16.47	17.88 (+8.6%)	19.11 (+16.0%)
	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%)
	SFT	17.67	18.89 (+6.9%)	20.42 (+15.6%)
	COPSD	18.12	21.18 (+16.9%)	23.55 (+30.0%)

Table 3: Average test-time scaling results across languages under generation budgets of 1,024, 2,048, and 4,096 tokens. Values in parentheses indicate the relative change compared with the corresponding 1,024-token budget. COPSD consistently achieves the strongest performance across model sizes and generation budgets

reasoning. To assess their individual contributions, we compare COPSD with two ablations: (i) the teacher receives only $x^{(H)}$, and (ii) the teacher receives only y^*. We train each variant using the same setup as in §5 and evaluate the resulting models on AfriMGSM. Figure 5 shows results.

Both the English problem and the reference solution are needed for the strongest supervision. Both ablated variants perform slightly worse than the full COPSD setting, suggesting that the two sources of privileged information play complementary roles: (i) When the teacher receives only the English problem, it may better interpret the low-resource input through a high-resource representation, but it lacks access to the correct solution and may therefore provide weaker guidance toward the desired reasoning trajectory. (ii) When the teacher receives only the reference solution, it has access to the target outcome, but may struggle to align that solution with the low-resource problem, creating a mismatch between the student input and the teacher context. Nevertheless, both variants still outperform the base model, indicating that each type of privileged information is useful on its own.

6.3 Test-Time Scaling

Larger models benefit more consistently from increased test-time computation. Figure 4 shows test-time scaling trends for three representative languages (Amharic, Ewe, and Zulu),[6] while Table 3 reports average results across all 17 low-resource AfriMGSM languages. Increasing the generation budget generally improves Pass@12, but the effect is clearer and more stable for larger models. For example, the Qwen3-8B base model improves from 14.73 at 1,024 tokens to 19.41 at 4,096 tokens, while COPSD improves from 18.12 to 23.55. By contrast, the gains for Qwen3-1.7B are smaller. This suggests that effective test-time scaling requires sufficient model capacity: larger models are better able to use additional generation budget to explore longer reasoning trajectories in low-resource languages, consistent with Yong et al. (2025).

COPSD strengthens the model's ability to use longer reasoning traces. Across all model sizes and generation budgets, COPSD achieves the highest average Pass@12. This indicates that the gains from COPSD persist as more test-time computa-

[6]We provide complete test-time scaling results in §C.

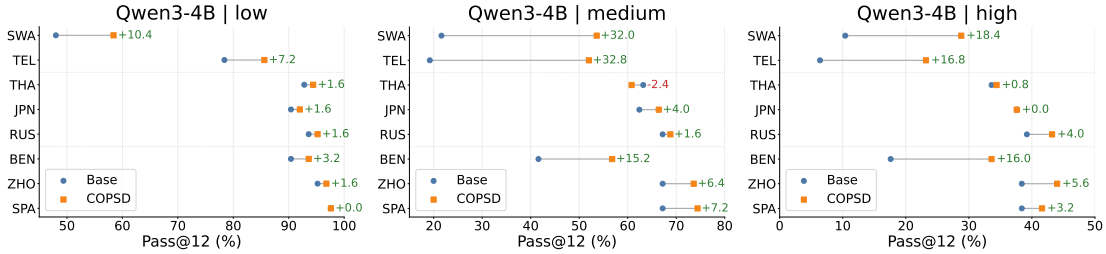

Figure 6: Pass@12 improvements of COPSD over the base model on PolyMath across low-, medium-, and high-difficulty settings for a 8,192-token generation budget. Each plot compares Base and COPSD for 8 languages spanning different resource levels. COPSD yields consistent gains across resource levels and difficulty levels, with especially substantial improvements for lower-resource languages such as Swahili (SWA) and Telugu (TEL).

tion is allocated. More importantly, COPSD often amplifies the benefit of longer generation budgets, especially for `Qwen3-8B`: its average performance increases by 30.0% from 1,024 to 4,096 tokens, compared with 13.8% for GRPO and 15.6% for SFT. Figure 4 provides concrete examples. For Amharic and Zulu with `Qwen3-8B`, COPSD starts close to the baselines at the 1,024-token budget, but separates more clearly as the budget increases. This pattern is particularly strong for Zulu, where COPSD reaches roughly 28% Pass@12 at 4,096 tokens, compared with about 16% for the baselines. These results suggest that COPSD improves not only low-resource reasoning performance, but also the model's ability to leverage longer target-language reasoning traces at inference time.

6.4 Generalization to Harder Benchmarks

To examine whether the gains from COPSD transfer beyond AfriMGSM, we further evaluate on **PolyMath** (Wang et al., 2025c), a more challenging multilingual mathematical reasoning benchmark with multiple difficulty levels. We select 8 languages spanning different resource levels, including low-resource languages: Swahili (SWA) and Telugu (TEL), mid- or high-resource languages: Thai (THA), Russian (RUS), Bengali (BEN), Japanese (JPN), Chinese (ZHO), and Spanish (SPA). We train `Qwen3-4B` with COPSD on each language using the same training setup as in §5, and evaluate on the low-, medium-, and high-difficulty subsets of PolyMath. For evaluation, we allow each model to generate up to 8,192 tokens. The results are shown in Figure 6.

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, indicating that the crosslingual reasoning behavior learned by COPSD is not limited to 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 substantially outperforms GRPO and SFT. Further analysis shows that COPSD improves format adherence, converges rapidly, strengthens test-time scaling, and generalizes to harder multilingual reasoning benchmarks. Together, these results suggest that COPSD offers an effective path to mitigate the difficulty of accessing and expressing latent reasoning ability through underrepresented languages.

Limitations

While COPSD consistently improves over the baselines across languages, several limitations remain and point to directions for future work.

First, COPSD uses English as the high-resource privileged language and assumes access to English reference solutions. This may limit its applicability in settings where high-quality English supervision is unavailable or where another high-resource language would provide a better reasoning signal.

Second, our training questions are translated from English into the target low-resource languages. Although COPSD does not require translated reasoning traces, translation artifacts in the problem statements may still affect training quality and downstream performance.

Finally, COPSD relies on the same model as the privileged teacher. When the model has limited competence in a target language, the teacher distribution may still be imperfect, even with access to English context and reference solutions. This may cause the learning signal to saturate quickly or degrade with continued training, as observed for some languages and model sizes.

Ethical Considerations

Use of AI Assistants. The authors used `ChatGPT` to assist with language polishing, including grammar, clarity, and coherence, as well as minor code implementation support.[7] 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, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. 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.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta. Association for Computational Linguistics.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. 2026. Long chain-of-thought reasoning across languages. *Preprint*, arXiv: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. 2024. Systematically analyzing prompt injection vulnerabilities in diverse llm architectures. *Preprint*, arXiv: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, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv: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 Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

Fahim Faisal, Kaiqiang Song, Song Wang, Simin Ma, Shujian Liu, Haoyun Deng, and Sathish Reddy Indurthi. 2025. Aligning multilingual reasoning with verifiable semantics from a high-resource expert model. *Preprint*, arXiv:2509.25543.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*,

[7] https://chatgpt.com/

pages 8920–8936, Suzhou, China. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. 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.

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, and 31 others. 2025. Openthoughts: Data recipes for reasoning models. *Preprint*, arXiv: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, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. 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.

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

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

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete, and Paul Pu Liang. 2025. Learn globally, speak locally: Bridging the gaps in multilingual reasoning. *Preprint*, arXiv:2507.05418.

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Geunbae Lee. 2026. Why do multilingual reasoning gaps emerge in reasoning language models? *Preprint*, arXiv:2510.27269.

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2026. What makes good multilingual reasoning? disentangling reasoning traces with measurable features. *Preprint*, arXiv:2604.04720.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. 2026. Why does self-distillation (sometimes) degrade the reasoning capability of llms? *Preprint*, arXiv:2603.24472.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. 2019. Spoc: Search-based pseudocode to code. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11883–11894.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Yihong Liu, Raoyuan Zhao, Hinrich Schütze, and Michael A. Hedderich. 2026. Large reasoning models are (not yet) multilingual latent reasoners. *Preprint*, arXiv:2601.02996.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *Preprint*, arXiv:2503.20783.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*.

Felix Petersen, Moritz Schubotz, Andre Greiner-Petter, and Bela Gipp. 2023. Neural machine translation for mathematical formulae. In *Proceedings of the 61st*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11534–11550, Toronto, Canada. Association for Computational Linguistics.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking language comes at the cost of accuracy. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20279–20296, Suzhou, China. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *Preprint*, arXiv:2404.04925.

Leonardo Ranaldi and Giulia Pucci. 2025. Multilingual reasoning via self-training. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11566–11582, Albuquerque, New Mexico. Association for Computational Linguistics.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. 2026. Crisp: Compressed reasoning via iterative self-policy distillation. *Preprint*, arXiv: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. 2023. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4945–4977, Singapore. Association for Computational Linguistics.

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

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

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

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Lintang Sutawika, Gokul Swamy, Zhiwei Steven Wu, and Graham Neubig. 2026. Gained in translation: Privileged pairwise judges enhance multilingual reasoning. *Preprint*, arXiv:2601.18722.

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

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Reasoning transfer for an extremely low-resource and endangered language: Bridging languages through sample-efficient language understanding. *Preprint*, arXiv: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. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schuetze. 2025a. Language mixing in reasoning language models: Patterns, impact, and internal causes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2637–2665, Suzhou, China. Association for Computational Linguistics.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025b. Demystifying multilingual reasoning in process reward modeling. In *Findings*

of the Association for Computational Linguistics: EMNLP 2025, pages 9775–9788, Suzhou, China. Association for Computational Linguistics.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *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*.

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. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *Preprint*, arXiv:2506.14245.

Linjuan Wu, Hao-Ran Wei, Baosong Yang, and Weiming Lu. 2025. From English to second language mastery: Enhancing LLMs with cross-lingual continued instruction tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23006–23023, Vienna, Austria. Association for Computational Linguistics.

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

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

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. 2026. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *Preprint*, arXiv:2602.12125.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.

Junjie Ye, Yuming Yang, Yang Nan, Shuo Li, Qi Zhang, Tao Gui, Xuanjing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. 2025. Analyzing the effects of supervised fine-tuning on model knowledge from token and parameter levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 471–513, Suzhou, China. Association for Computational Linguistics.

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

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. 2026a. Opsdl: On-policy self-distillation for long-context language models. *Preprint*, arXiv:2604.17535.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2026b. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning. *Preprint*, arXiv:2510.07300.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *Preprint*, arXiv:2401.01055.

Raoyuan Zhao, Yihong Liu, Hinrich Schuetze, and Michael A. Hedderich. 2026a. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5223–5247, Rabat, Morocco. Association for Computational Linguistics.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026b. Self-distilled reasoner: On-policy self-distillation for large language models. *Preprint*, arXiv:2601.18734.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

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 ƙarshe 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	ꕉ ꔖꔤꕞ꘡ ꕢ ꕊꔤ ꕞꕞ ꕞ ꕤ ꔖꔤ꘡ ꕢ ꕥꔤꕞ ꕉꕞꕞꕤꕢ ꕉ ꕤ ꕉꔤꕞꕤ ꕤ \boxed{}.
WOL	Ba beneen yoon, xalaatal ci ndànk-ndànk, te defal sa tontu mu mujj mi ci \boxed{}.
YOR	Jọ̀wọ́ ronú ní ìgbésẹ̀-ní-ìgbésẹ̀, kí o sì fi ìdáhùn ìkẹyìn rẹ sínú \boxed{}.
AMH	እባክዎ ደረጃ በደረጃ ያስቡ፣ የመጨረሻ መልስዎንም በ \boxed{} ውስጥ ያስቀምጡ።
KIN	Nyamuneka tekereza intambwe ku yindi, kandi ushyire igisubizo cya nyuma muri \boxed{}.
LUG	Nsaba lowooza mutendera ku mutendera, era oteeke eky'okuddamu ekisembayo mu \boxed{}.
SWA	Tafadhali fikiri hatua kwa hatua, na uweke jibu lako la mwisho ndani ya \boxed{}.
ORM	Maaloo tartiiba tartiibaan yaadi, deebii kee isa dhumaa \boxed{} keessa kaa'i.
SNA	Ndapota funga nhanho nenhanho, uye isa mhinduro yako yekupedzisira mukati me \boxed{}.
XHO	Nceda ucinge inyathelo ngenyathelo, uze ufake impendulo yakho yokugqibela ngaphakathi kwe \boxed{}.
ZUL	Sicela ucabange isinyathelo ngesinyathelo, bese ufaka impendulo yakho yokugcina ngaphakathi kwe \boxed{}.
SOT	Ka kopo nahana mohato ka mohato, 'me u kenye karabo ea hao ea ho qetela ka hare ho \boxed{}.
LIN	Nabondeli yo, kanisá litambe na litambe, mpe tyá eyano na yo ya nsuka na kati ya \boxed{}.

Figure 7: Instructions used for each target-language math problem to encourage step-by-step reasoning and require the final answer to be placed inside \boxed{}.

Lang.	Prefix
EWE	Le biabia me la, magɔme le susu wɔwɔ me le Eʊegbe 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	Eꕀ ꕉ ꕊꔤꕢ ꕤ꘡ ꕢ ꔤꕞ ꕥꔤꕞ ꕉ ꕤ ꕉꕢ ꕤ꘡
WOL	Ci laaj bi, dinaa tàmbali xalaat ci Wolof.
YOR	Gẹ́gẹ́ bí a ti béèrè, 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.

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

Table 4: Language coverage of our experiments. We use the ISO 639-3 codes as language identifiers.

A Experimental Details

A.1 Language coverage

Our experiments cover all 17 African languages included in AfriMGSM (Adelani et al., 2025). For each language, we use its ISO 639-3 code as the language identifier throughout training, evaluation, and result reporting. The covered languages span multiple language families and writing systems. This setting allows us to evaluate whether COPSD can improve reasoning not only across different languages, but also across substantially different orthographic and linguistic conditions. Table 4 lists the languages, ISO 639-3 codes, and target-language names used in our experiments.

A.2 Language Control

Following Qi et al. (2025); Zhao et al. (2026a), we use complementary prompting strategies to encourage the model to produce its explicit reasoning trace in the target low-resource language.

Language-Specific Instruction For each input, we prepend a language-specific instruction that specifies the desired reasoning language and asks the model to solve the problem step by step. The language-specific instructions are shown in Figure 7.

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.

B Qualitative Analysis of Reasoning Trace

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.

COPSD effectively mitigates repetitive degeneration in reasoning traces. Figure 9 reports the average 4-gram repeat rate of `Qwen-1.7B` throughout training.[8] 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

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

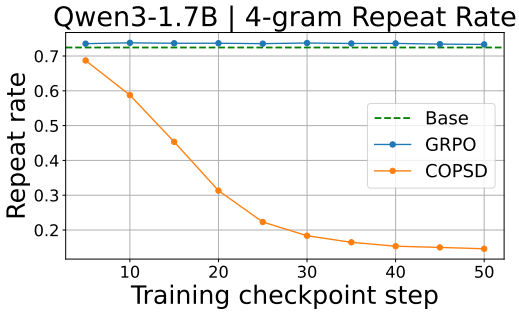

Figure 9: Average repeat rate comparison on `Qwen3-1.7B` with 4-grams. COPSD consistently reduces repetition compared to the base model and GRPO.

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.

C 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 3 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 $n = 1$ 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.

D 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.[9]

D.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.[10]

D.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.

D.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

[9] Hidden for anonymity.

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

templates for every language in our experiments are provided in our GitHub repository.

E 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 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 and Ba, 2015; Loshchilov and 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 5 summarizes GRPO and COPSD hyperparameters, while Table 6 reports the SFT configuration.

For evaluation, we use the same decoding configuration for all models and methods to ensure fair comparison, as shown in Table 7. 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 the benchmark consists of relatively short mathematical reasoning problems. Final answers are extracted from `\boxed{}` and verified as described in §5. For each language and method, we select the checkpoint that achieves the best performance under the 1,024-token budget, and then report that checkpoint's performance under the other generation budgets.

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

Table 5: 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	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

Table 6: Training hyperparameters for the SFT baseline on self-generated reasoning traces. SFT uses the concatenation of the student prompt and generated teacher completion as the training sequence, with a maximum sequence length of 8192 tokens.

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

Table 7: Inference hyperparameters used for evaluation. We use the same decoding configuration for all models and methods.

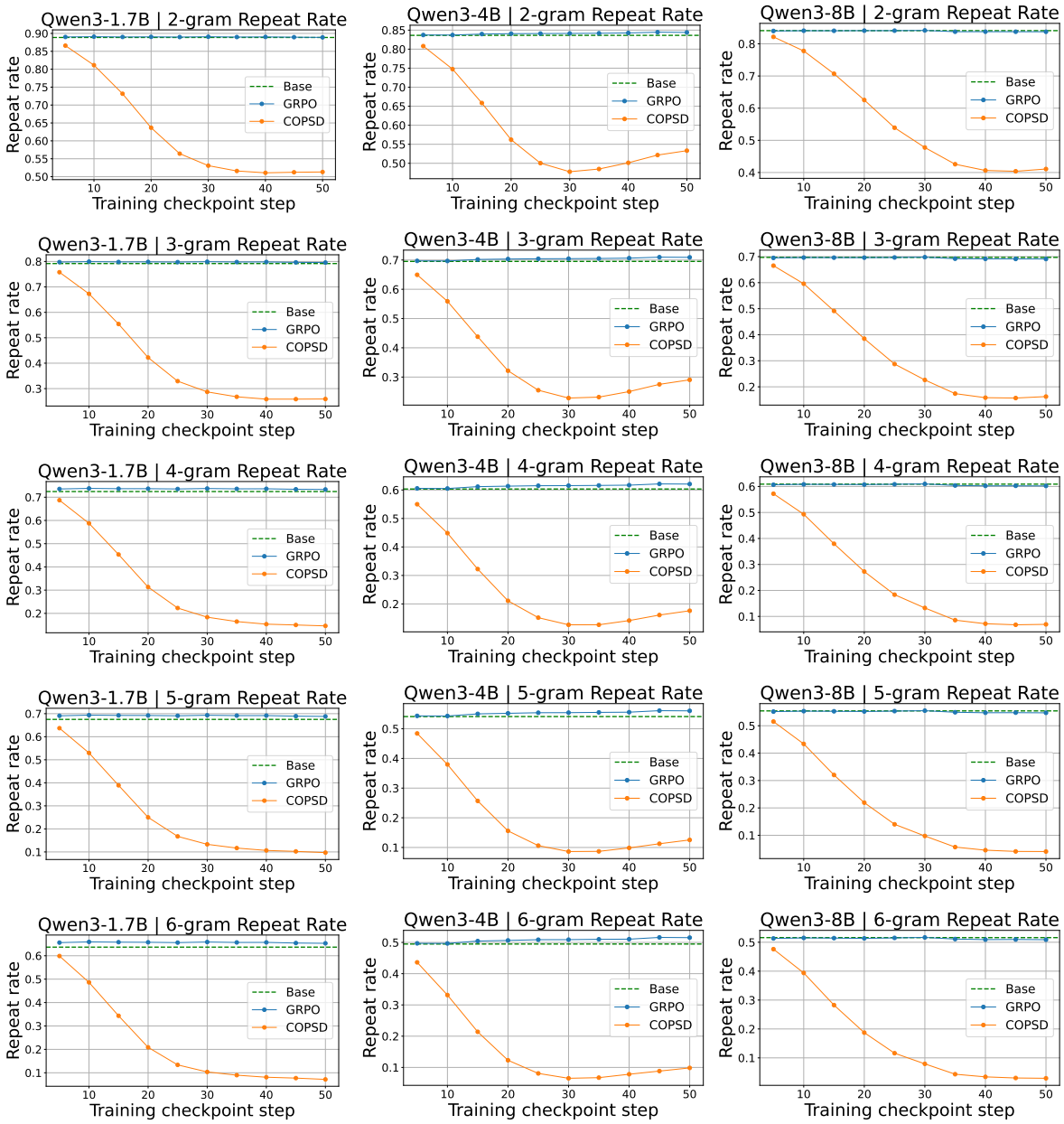

Figure 10: Repeat rate across different n-gram settings ($n = 2$ to 6) and model sizes. COPSD consistently reduces repetition compared to both the base and GRPO.

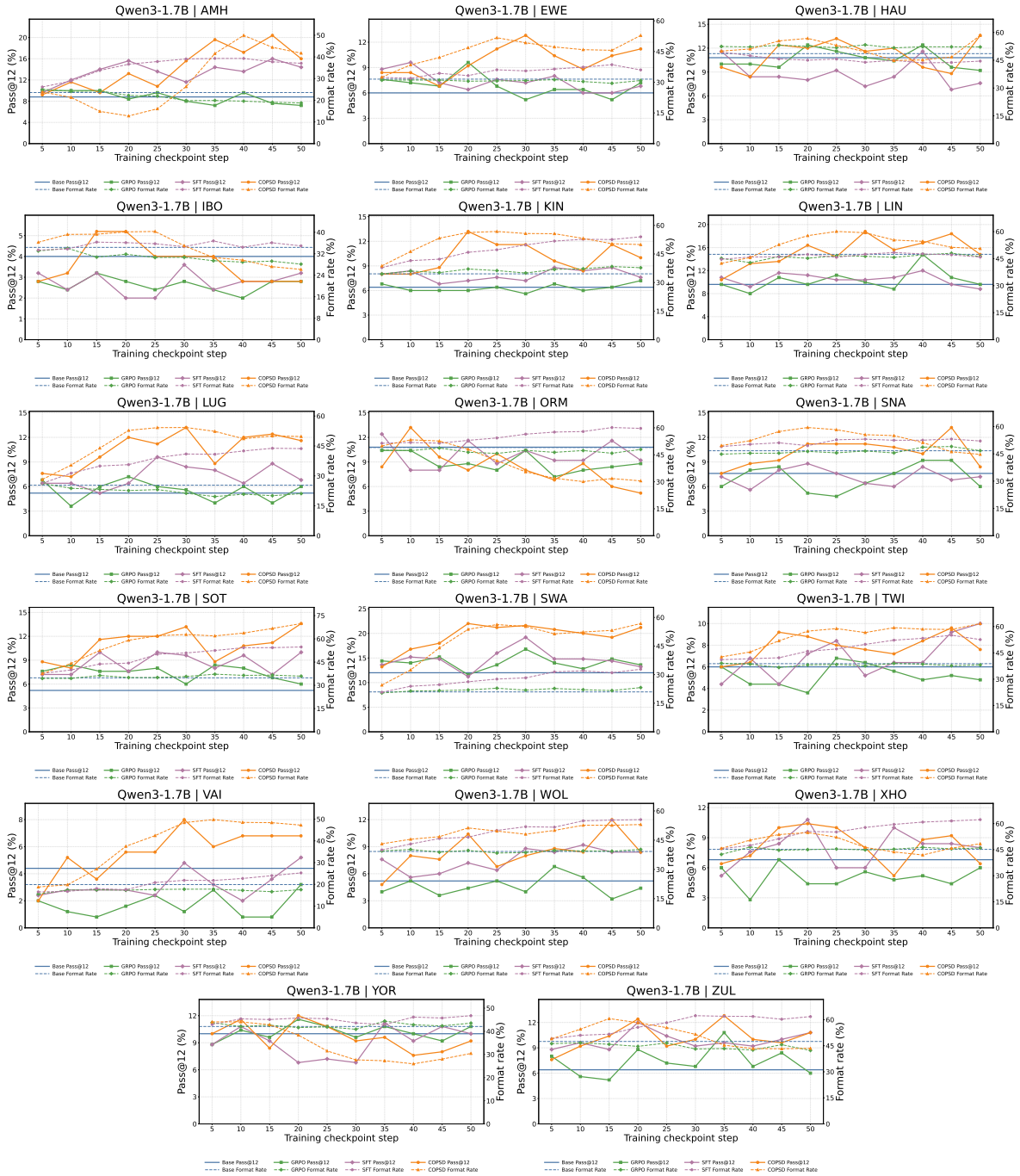

Figure 11: Per-language training dynamics for `Qwen3-1.7B` across all African languages under a 1024-token generation budget. Solid lines show Pass@12 and dashed lines show format rate.

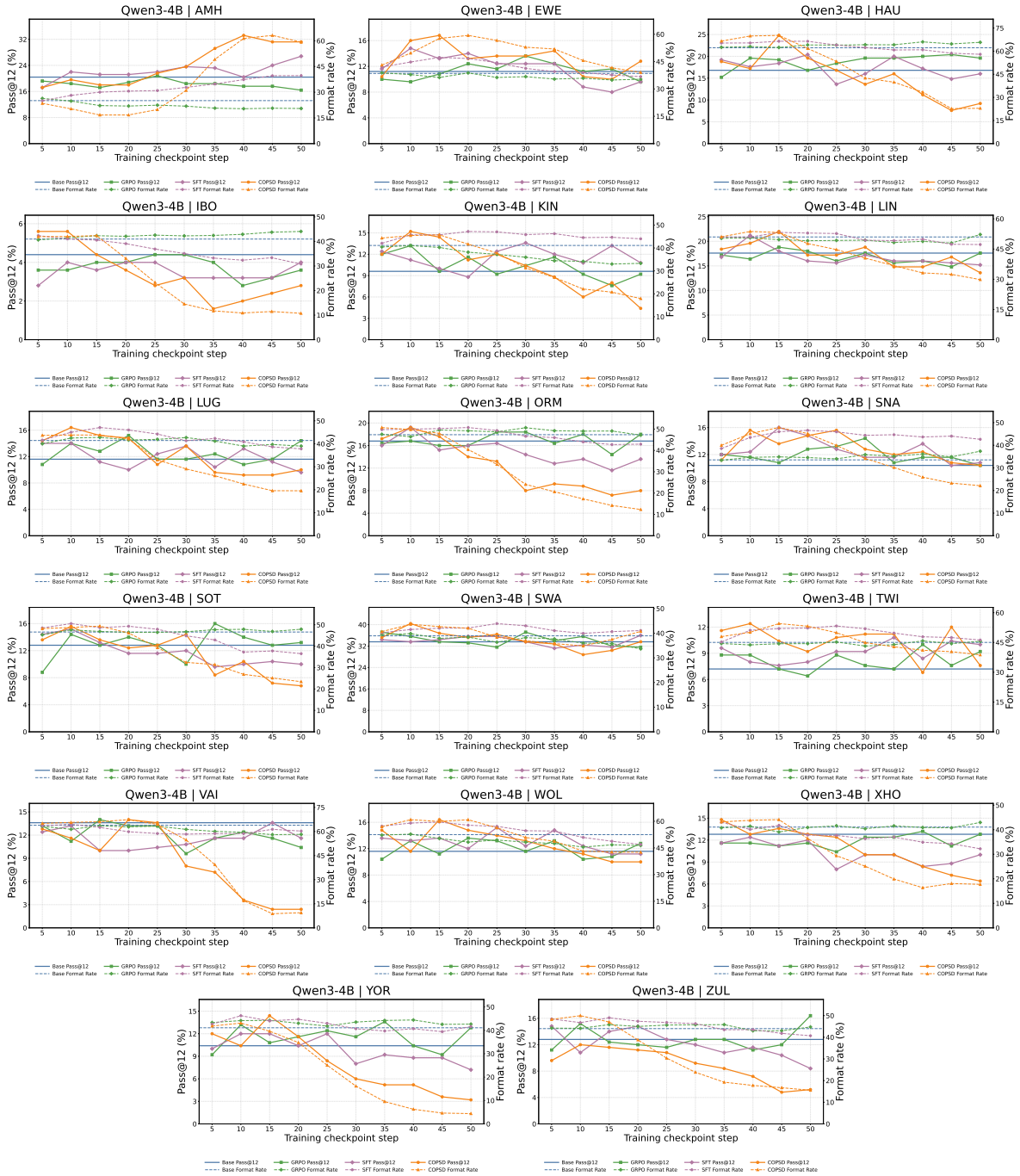

Figure 12: Per-language training dynamics for `Qwen3-4B` across all African languages under a 1024-token generation budget. Solid lines show Pass@12 and dashed lines show format rate.

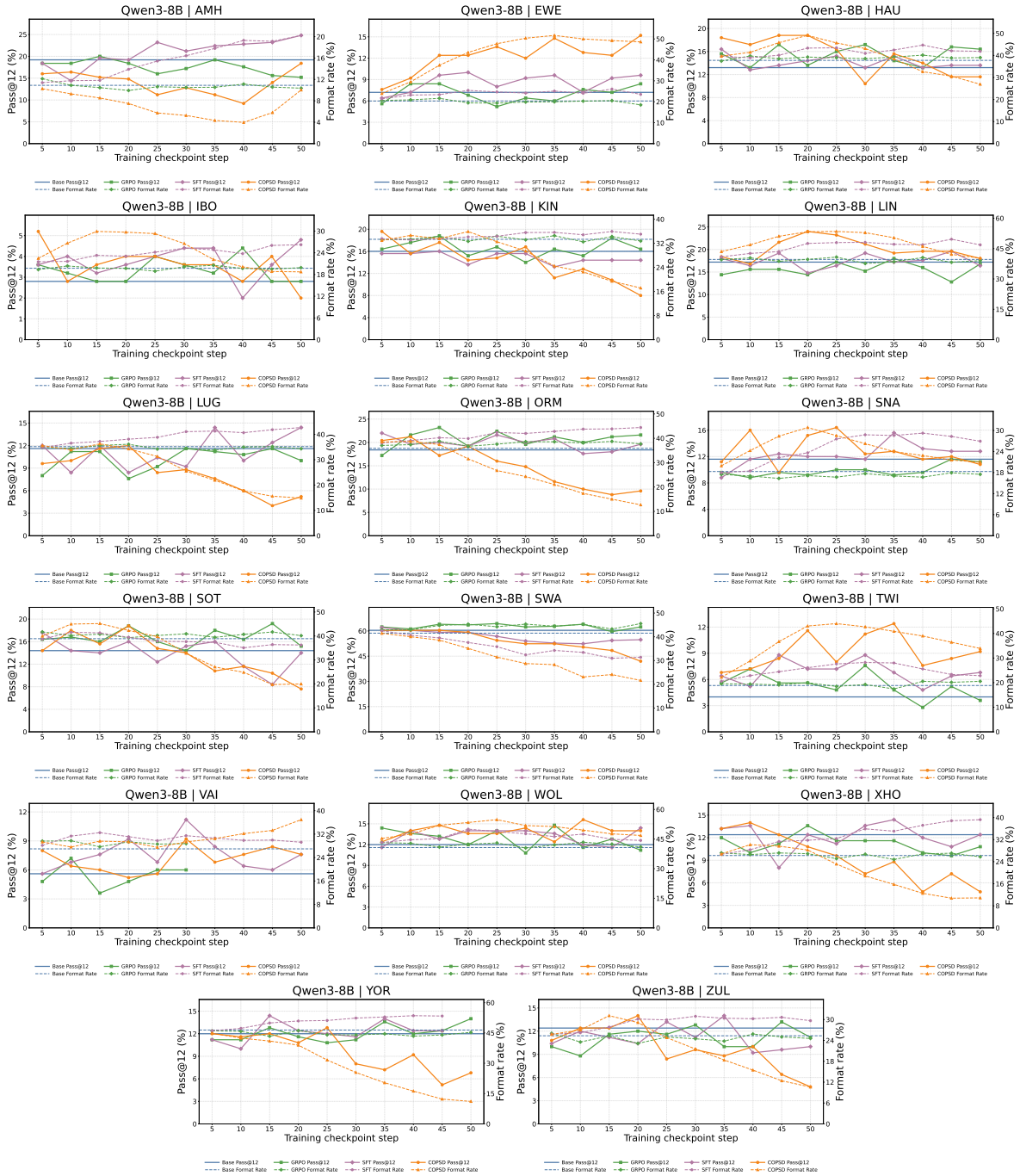

Figure 13: Per-language training dynamics for `Qwen3-8B` across all African languages under a 1024-token generation budget. Solid lines show Pass@12 and dashed lines show format rate.

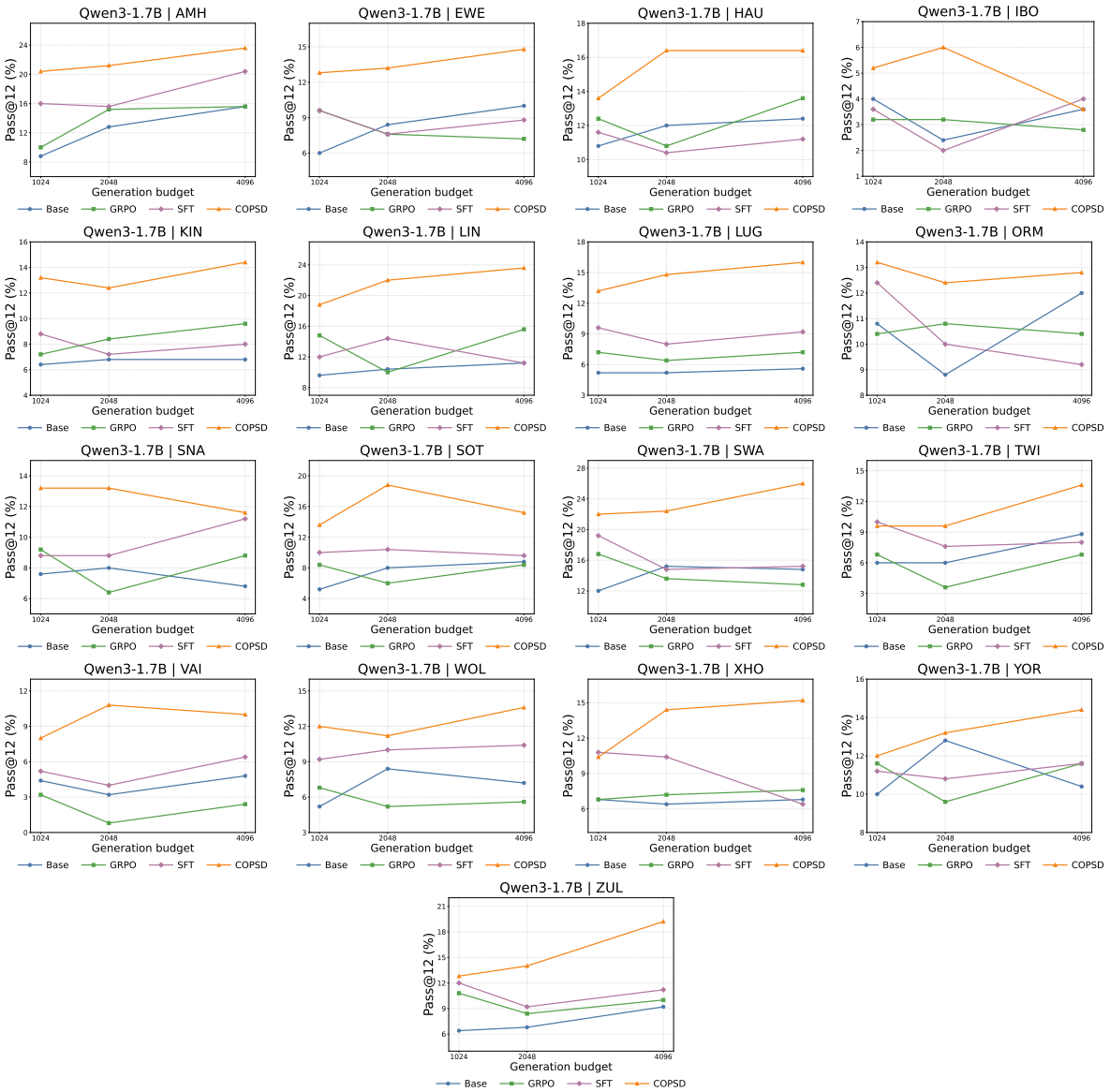

Figure 14: Per-language test-time scaling results on Pass@12 for `Qwen3-1.7B` across all African languages, under generation budgets of 1024, 2048, and 4096 tokens. Overall, the trends are mixed across languages, but COPSD generally achieves stronger performance than both the Base model and GRPO under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

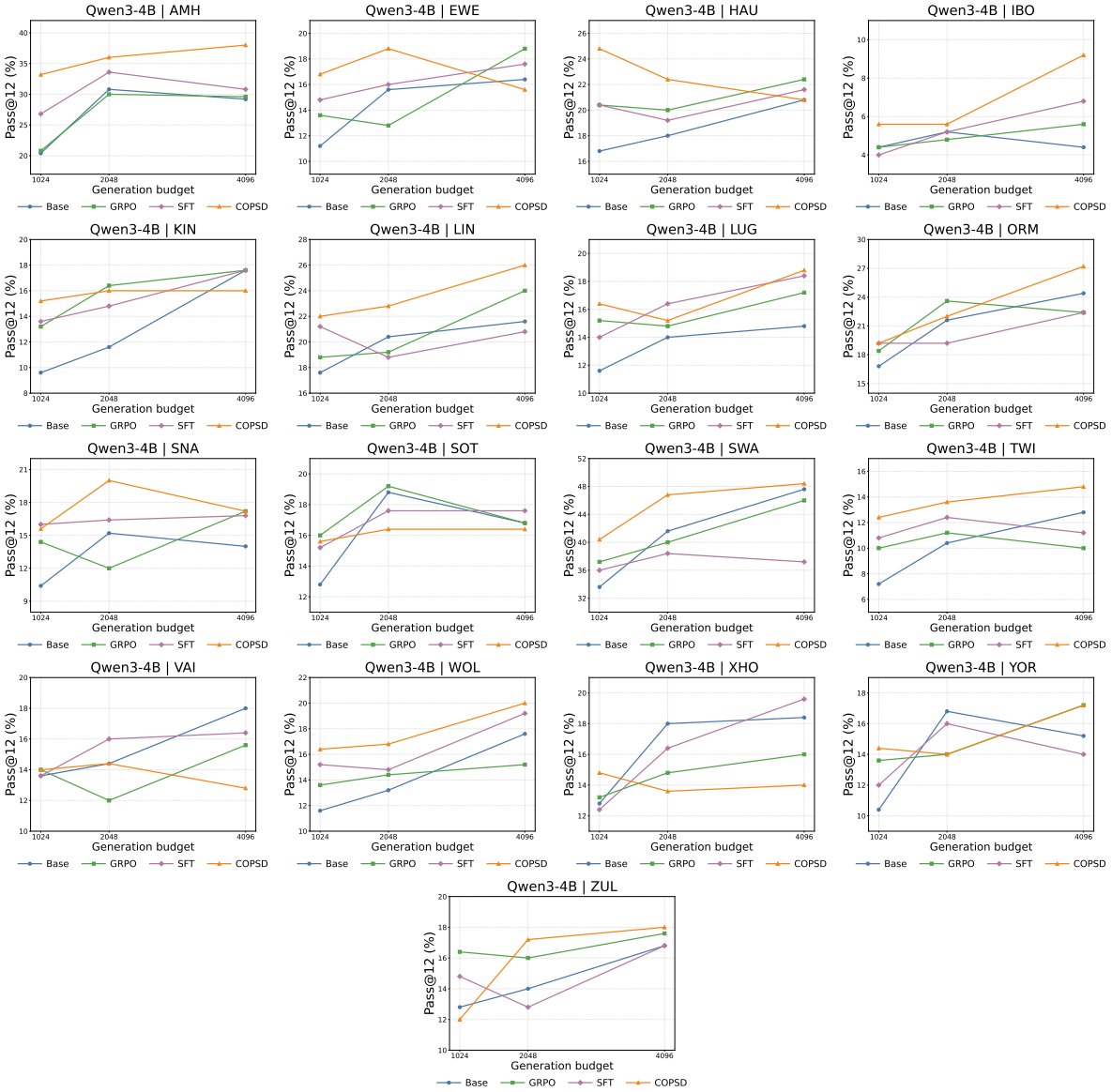

Figure 15: Per-language test-time scaling results on Pass@12 for `Qwen3-4B` across all African languages, under generation budgets of 1024, 2048, and 4096 tokens. Overall, the trends are mixed across languages, but COPSD generally achieves stronger performance than both the Base model and GRPO under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

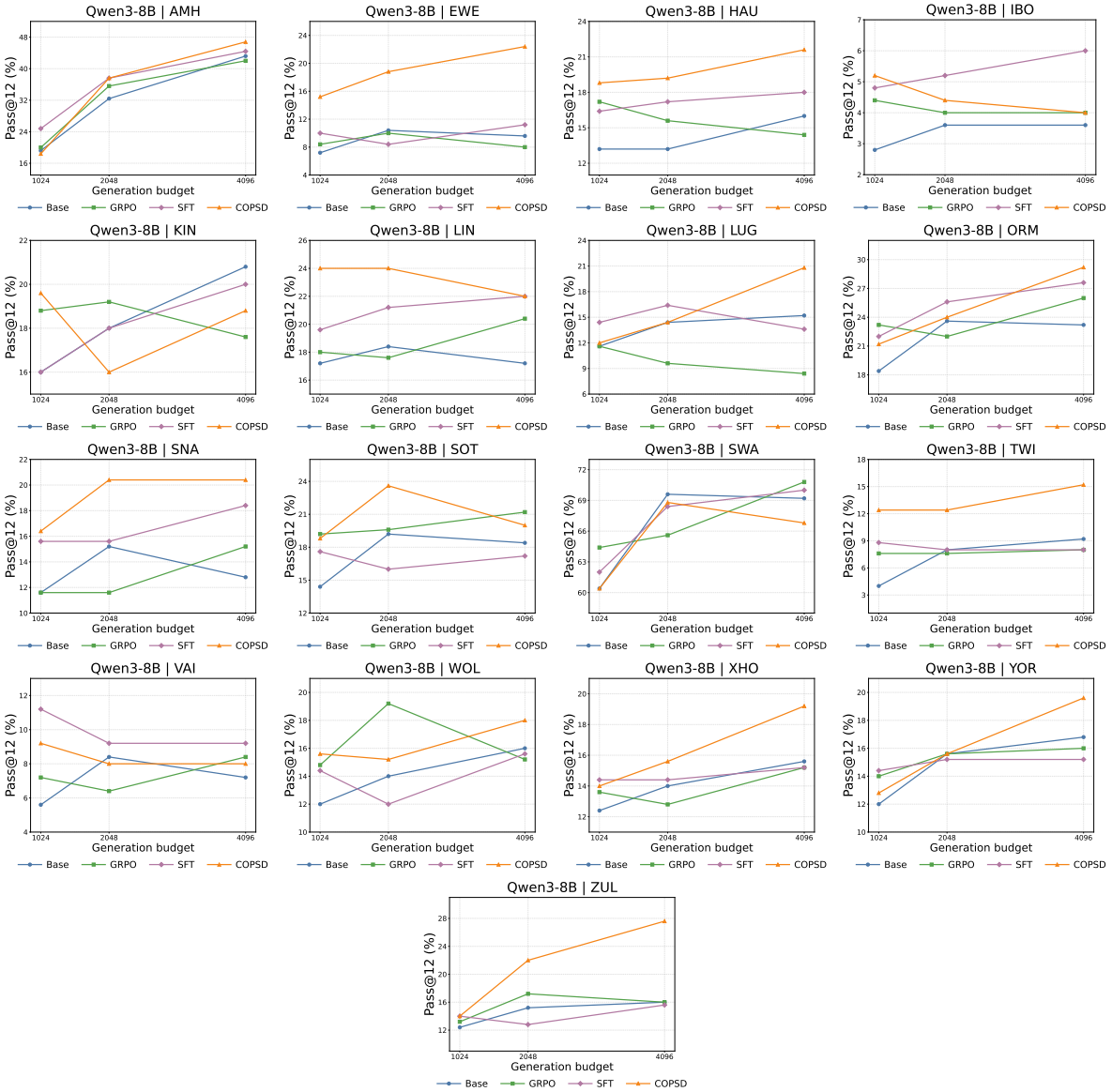

Figure 16: Per-language test-time scaling results on Pass@12 for `Qwen3-8B` across all African languages, under generation budgets of 1024, 2048, and 4096 tokens. Overall, the trends are mixed across languages, but COPSD generally achieves stronger performance than both the Base model and GRPO under 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)	English reference translation
Swali: *[problem_target]* *Tafadhali fikiri hatua kwa hatua, na uweke jibu lako la mwisho ndani ya* `\boxed{}`.	**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)	English reference translation

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{}`.
	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.