CROSS-LINGUAL MULTI-VIEW ON-POLICY SELF-DISTILLATION: DISTILLING MULTILINGUAL VIEWS INTO A SINGLE TARGET-LANGUAGE POLICY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual language models often exhibit weaker reasoning capabilities in low-resource languages. Cross-lingual reasoning can leverage high-resource-language renderings of the same problem at inference time, but doing so requires repeated multilingual decoding and answer aggregation for every query. We ask whether this inference-time benefit can instead be internalized into a model that reasons directly in the target language. We introduce **Cross-lingual Multi-view On-Policy Self-Distillation (CMOPD)**, which uses multiple language versions as privileged training-time views of the same problem. The views provide dense token-level feedback on one low-resource on-policy trajectory, and on-policy distillation consolidates this feedback into a single target-language policy. After training, the model answers directly from the low-resource query without multilingual decoding or answer aggregation. Across Qwen3 $\{$1.7B, 4B, 8B$\}$ backbones and four AfriMGSM target languages (Swahili, Igbo, Yoruba, Amharic), CMOPD with English and Chinese views improves mean pass@12 over our reproduced single-view baseline at every scale (1.7B: $+0.2$, 4B: $+2.6$, 8B: $+1.8$), while per-language effects remain heterogeneous. On Qwen3-8B Swahili, adding the Chinese view improves training-time distillation by 4.8 points; multilingual test-time synthesis improves over English-only reasoning by 4.4 points.

1 INTRODUCTION

Large language models (LLMs) have advanced rapidly in recent years, with substantial improvements in language understanding, knowledge-intensive tasks, and complex reasoning (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023). However, training data are unevenly distributed across languages. Low-resource languages have far less training data than high-resource ones (Üstün et al., 2024). This data gap often results in weaker model capabilities in low-resource languages, particularly for complex reasoning (Adelani et al., 2024). This leaves their speakers at a disadvantage compared with users of well-resourced languages. Improving LLM capabilities in low-resource languages is therefore an important research problem.

Prior work has transferred supervision from high-resource languages through supervised fine-tuning (SFT) or reinforcement learning (RL). Cross-lingual SFT trains models on translated or synthesized reasoning traces (Upadhayay & Behzadan, 2023; Chai et al., 2024; Zheng et al., 2025; Zhang et al., 2026), but often favors imitation over generalization to unseen problems (Chu et al., 2025; Kapusuzoglu et al., 2026). Cross-lingual RL instead learns from model-generated responses and outcome rewards, but typically requires multiple costly rollouts for each problem (Faisal et al., 2025). Recent work applies on-policy distillation (OPD) to cross-lingual capability transfer. In particular, Cross-lingual On-Policy Self-Distillation (COPSD) (Liu et al., 2026) uses an English-conditioned teacher to provide token-level feedback on trajectories sampled from the target-language policy. However, COPSD fixes the teacher to English, limiting its supervision to a single reasoning language.

Cross-lingual chain-of-thought (cross-lingual CoT) suggests why this may be limiting: combining or adaptively selecting reasoning languages can outperform relying on a single language (Qin et al.,

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.

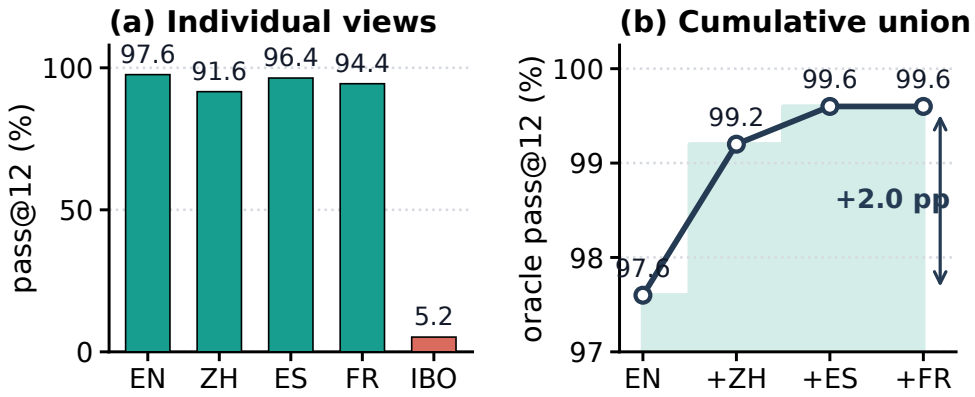


Figure 1: **Different language views yield different success sets.** Qwen3-1.7B is evaluated on the same 250 AfriMGSM-Igbo problems with 12 samples per language rendering. **(a)** Pass@12 for each view. **(b)** The oracle union rises from 97.6 with English to 99.6 across English, Chinese, Spanish, and French, motivating whether multi-view behavior can be internalized during training.

2023; Zheng et al., 2025; Yu et al., 2025). Prior work also finds stronger transfer between more closely related languages (Muller et al., 2022), suggesting that the choice of language views may matter. Figure 1 illustrates this complementarity: different language views solve overlapping but non-identical sets of problems. Combining multiple reasoning languages at inference requires repeated generation and answer aggregation for every query. This leads us to ask: *Can multilingual test-time computation be internalized into a single target-language policy?*

We answer this question with **Cross-lingual Multi-view On-Policy Self-Distillation (CMOPD)**, which distills multiple high-resource-language views into a single target-language policy. CMOPD samples one trajectory from the target-language policy. A shared frozen backbone evaluates each next token under several translated problem–solution views, and their predictions are combined into a teacher target. Multilingual computation is confined to training; after training, the model answers directly in the target language with a single response (§3).

Our contributions are threefold. **(i)** We formulate multilingual reasoning views as privileged training signals for a target-language policy. **(ii)** We introduce CMOPD, which internalizes these views through a frozen self-teacher on trajectories sampled from the target-language policy. **(iii)** Across three Qwen3 scales and four African target languages, CMOPD improves mean pass@12 over reproduced COPSD at every scale; our analyses characterize cross-view disagreement and target fusion (§4).

2 Related Work

Multilingual reasoning and capability transfer. Multilingual reasoning studies whether language models can solve reasoning problems expressed in different languages. A central challenge is transferring capabilities from better-supported languages to lower-resource ones while enabling the model to reason directly in the target language. Existing approaches use translated or multilingual supervision for SFT (Chai et al., 2024; Zhang et al., 2026), align representations across languages (Yoon et al., 2024), or optimize the target-language policy through RL (Faisal et al., 2025).

Cross-lingual reasoning and distillation. A more specific line of work uses a high-resource language as an intermediate reasoning view. Cross-lingual chain-of-thought methods translate the problem or prompt the model to reason in another language (Huang et al., 2023; Qin et al., 2023). Later methods adaptively select reasoning languages or aggregate predictions across multilingual reasoning paths (Zheng et al., 2025; Yu et al., 2025). These approaches exploit language-conditioned reasoning through inference-time generation and answer-level selection or aggregation. COPSD instead brings cross-lingual supervision into training: an English-conditioned teacher provides token-level feedback on trajectories sampled from the target-language policy (Liu et al., 2026). CMOPD builds

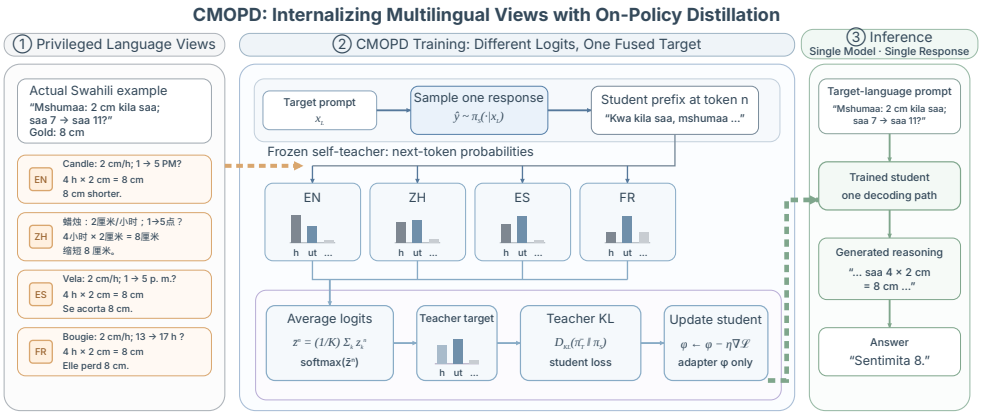


Figure 2: **Internalizing multilingual reasoning through OPD.** A low-resource-language problem is translated into several high-resource languages, producing multiple language-conditioned problem–answer views. The target-language policy samples one reasoning trajectory, and the shared teacher backbone uses each multilingual view to predict the next token along that trajectory. CMOPD combines these predictions and self-distills them into the target-language policy. Inference uses the original low-resource-language problem, one model, and one response.

on this setting but uses multiple language-conditioned teacher views rather than a single English view.

Multi-teacher on-policy distillation. Recent work extends OPD from a single teacher to multiple specialized teachers. The MiMo-V2-Flash technical report introduced Multi-Teacher On-Policy Distillation (MOPD), in which prompts are routed to domain-specific teachers that supervise student-generated trajectories (Xiaomi LLM-Core Team, 2026). The subsequent MOPD study formalizes this approach as a general framework for integrating independently trained domain capabilities into one student (Ma et al., 2026). Nemotron-Cascade 2 similarly uses the strongest intermediate teacher for each domain to recover capabilities during multi-stage post-training (Yang et al., 2026). MAD-OPD instead lets multiple teachers debate over the same on-policy state and combines their predictions using post-debate confidence (Wang et al., 2026). These methods obtain complementary supervision from distinct teacher models or checkpoints. CMOPD instead uses multiple language-conditioned views of the same frozen model on a shared target-language trajectory, introducing no additional teacher models.

3 Method

3.1 Method Overview

CMOPD uses multilingual reasoning views as privileged training signals and internalizes their complementary supervision into a target-language policy. Each view consists of the same problem and reference solution translated into a high-resource language and is available only during training. We implement this internalization through on-policy self-distillation, so that the trained policy can reason directly from the original target-language problem without translated views at inference time.

Figure 2 summarizes the training procedure. Given a target-language problem, the student first samples a reasoning trajectory from its current policy. A frozen self-teacher evaluates this trajectory under each privileged high-resource-language view and produces a next-token distribution at every position. CMOPD combines the distributions from all views into a single teacher target and trains the student on the same on-policy trajectory.

3.2 LANGUAGE-CONDITIONED TEACHER CONTEXTS

Let x_L denote a problem in a low-resource target language L. We translate the problem and its reference solution into K high-resource languages. For each $k \in \{1, \dots, K\}$, let H_k denote the k-th high-resource language. The translated problem x_{H_k} and reference solution $y^\star_{H_k}$ form the privileged context

$$c_k \triangleq (x_{H_k}, y^\star_{H_k}). \tag{1}$$

The contexts $\{c_k\}_{k=1}^K$ are available only to the frozen self-teacher during training. Each context provides the frozen self-teacher with a different language view of the same problem and reference solution.

3.3 ON-POLICY INTERNALIZATION OF MULTILINGUAL VIEWS

Let p_θ denote the frozen pretrained backbone and ϕ the trainable student adapter. The resulting student policy is

$$\pi_S(\cdot \mid x_L) \triangleq p_{\theta,\phi}(\cdot \mid x_L). \tag{2}$$

The student samples a target-language reasoning trajectory $\hat{y} \sim \pi_S(\cdot \mid x_L)$. For each position $n \in \{1, \dots, |\hat{y}|\}$, let π_S^n denote its next-token distribution conditioned on the prefix $\hat{y}_{<n}$.

The frozen self-teacher then evaluates the same trajectory under every privileged context. At position n, its next-token distribution under context c_k is

$$\pi_T^{(k),n}(\cdot) \triangleq p_\theta(\cdot \mid x_L, c_k, \hat{y}_{<n}). \tag{3}$$

All K views use the same frozen self-teacher and evaluate the same prefix $\hat{y}_{<n}$. Their next-token distributions can therefore be compared and combined position by position; they do not come from independently sampled target-language trajectories.

At each position, an aggregation operator $\mathcal{A}$ represents the joint supervision from all language views as one teacher target:

$$\bar{\pi}_T^n = \mathcal{A}\Big(\{\pi_T^{(k),n}\}_{k=1}^K\Big). \tag{4}$$

CMOPD internalizes this supervision by training the student against the per-token targets on its own trajectory:

$$\mathcal{L}_{\text{CMOPD}}(\phi) = \mathbb{E}_{\hat{y}\sim\pi_S}\left[\frac{1}{|\hat{y}|}\sum_{n=1}^{|\hat{y}|} D_{\text{KL}}(\bar{\pi}_T^n \,\|\, \pi_S^n)\right]. \tag{5}$$

Because $\hat{y}$ is sampled from the current policy, the multilingual supervision is applied at the states visited by the student itself. The teacher-view predictions are computed without gradients, and gradients flow only through π_S to the adapter parameters ϕ.

The explicit ordering shows that the implemented divergence is $D_{\text{KL}}(\bar{\pi}_T^n \| \pi_S^n)$. Following COPSD, the implementation clips divergence contributions for stability; the threshold is reported in the supplementary material. When $K{=}1$ and the teacher view is English, CMOPD reduces to COPSD.

3.4 VIEW AGGREGATION

CMOPD evaluates all K teacher views on the same student trajectory and combines their predictions within each update. Let $\mathbf{z}_k^n$ denote the teacher logits from view k at token position n. We average these logits and normalize the result to obtain the distillation target:

$$\bar{\pi}_T^n = \text{softmax}\left(\frac{1}{K}\sum_{k=1}^K \mathbf{z}_k^n\right). \tag{6}$$

This is equivalent to the normalized geometric mean of the per-view distributions. Section 4.3 compares this simultaneous fusion with sequential use of the views and with alternative choices of the fusion target.

Method	Qwen3-1.7B					Qwen3-4B					Qwen3-8B				
	SWA	IBO	YOR	AMH	Avg.	SWA	IBO	YOR	AMH	Avg.	SWA	IBO	YOR	AMH	Avg.
Base	14.80	3.60	10.40	15.60	11.10	47.60	4.40	15.20	29.20	24.10	69.20	3.60	16.80	43.20	33.20
GRPO†	12.80	2.80	11.60	15.60	10.70	46.00	5.60	17.20	29.60	24.60	70.80	4.00	16.00	42.00	33.20
COPSD (K=1, EN)	**29.60**	7.60	**19.60**	25.60	20.60	54.00	10.40	16.80	31.20	28.10	74.40	**11.60**	21.20	**49.60**	39.20
CMOPD (K=2, EN+ZH)	29.20	**9.20**	17.60	**27.20**	**20.80**	**56.40**	**11.20**	**18.00**	**37.20**	**30.70**	**79.20**	11.20	**26.40**	47.20	**41.00**

Table 1: **Main results on AfriMGSM pass@12.** Results use n=12 samples and `max_new_tokens`= 4096. Base and GRPO† are copied from Liu et al. (2026); COPSD and CMOPD are our matched runs. Bold marks the better of COPSD and CMOPD in each model/language column; the shaded row denotes CMOPD with EN+ZH views.

Inference. The multilingual views and the aggregation operation are used only for training. The trained model answers the original low-resource-language problem directly, with no translation, multilingual decoding, or answer voting.

4 Experiments

4.1 Setup

Models and data. We evaluate Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (Yang et al., 2025). For each model, the self-teacher is frozen and only a Low-Rank Adaptation (LoRA) adapter is trained. We use COPSD-AfriMGSM (Liu et al., 2026) for training, which contains 500 problems per language. Each problem and reference solution is additionally translated into Chinese, Spanish, and French using Qwen3-4B. We evaluate on the `masakhane/afrimgsm` test split, which contains 250 problems per language in Swahili (SWA), Igbo (IBO), Yoruba (YOR), and Amharic (AMH).

Baselines and evaluation. We compare CMOPD with the base model, Group Relative Policy Optimization (GRPO), and single-view COPSD (Liu et al., 2026). GRPO results are taken from Liu et al. (2026, Table 1). We exclude SFT because our teacher-forced runs produce English rather than target-language reasoning. We report pass@12 using 12 completions per problem, sampled with temperature 1.0, top-p 0.95, and a maximum generation length of 4096 tokens. Each result is obtained from a single trained checkpoint.

Training details. We train for 100 steps with a learning rate of 5×10^{-6}, an effective batch size of 32, and LoRA rank 64 with α=128. The maximum completion length is 2048, the sampling temperature is 1.1, and β is set to 0. Training uses a fixed teacher and a target-language thinking prefix. All runs use the Hugging Face default seed of 42 and eight NVIDIA A100 GPUs with colocated vLLM rollout. The supplementary material provides the full configuration and software environment.

4.2 Main Results

Table 1 reports pass@12 across four target languages and three Qwen3 model scales.

Comparison with reported Base and GRPO results. Across the three model scales, CMOPD improves average pass@12 over the reported base-model results by 9.7, 6.6, and 7.8 points, respectively. The corresponding differences from the reported GRPO results are 10.1, 6.1, and 7.8 points. The matched COPSD comparison isolates the contribution of adding the Chinese teacher view under the same training setup.

CMOPD improves average performance across model scales. Compared with our reproduced COPSD baseline, CMOPD improves average pass@12 by 0.2, 2.6, and 1.8 points on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. The gain is modest at 1.7B but becomes more pronounced at the two larger scales. CMOPD outperforms COPSD in eight of the twelve model–language comparisons, with the largest gains on 4B AMH (+6.0), 8B YOR (+5.2), and 8B SWA (+4.8). Output-

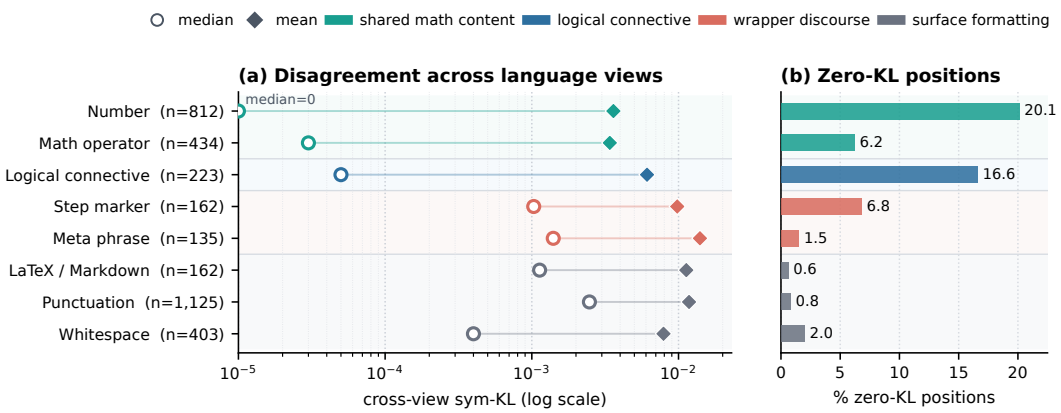


Figure 3: **Cross-view disagreement concentrates in wrapper-dependent tokens.** Pairwise symmetric KL is computed over English, Chinese, Spanish, and French teacher contexts on 15,360 Qwen3-8B-Instruct Swahili rollout positions. **(a)** Median and mean divergence. **(b)** Positions at which all four views produce identical next-token distributions.

format compliance remains above 91.7%, and average output length stays within 8% of COPSD across all CMOPD variants.

4.3 Multi-View Combination Ablation

Data-level combination. Sequential OPD (seqOPD) combines language views across training updates rather than within a teacher target. It applies the standard COPSD objective (Liu et al., 2026) using one teacher view for each optimizer update. We consider two schedules. The alternating schedule switches between EN and ZH after every update. The block schedule uses ZH for the first 50 updates and EN for the remaining 50. In both cases, all examples within an effective batch use the same teacher view; views are not assigned separately to individual queries.

Operator	Teacher-view usage	pass@12
COPSD	Train with the English teacher only	74.4
Sequential	Alternate English and Chinese teachers across updates, using one teacher at a time	75.6
Sequential	Train with the Chinese teacher for the first 50 updates and the English teacher for the remaining 50 updates	72.8
CMOPD	Fuse the English and Chinese teachers within every update	**79.2**

Table 2: Update-operator ablation on Qwen3-8B SWA. All entries are evaluated once after 100 training updates. Sequential OPD uses one teacher view per update; CMOPD uses parallel logit-mean fusion.

Alternating EN and ZH updates slightly improves over single-view COPSD, whereas the block schedule does not. Both remain below parallel EN+ZH fusion. We next compare several ways of combining the two views within each update.

Fusion rule. Fig. 4a compares five ways of using the EN and ZH teacher predictions on Qwen3-1.7B SWA. Logit mean averages the two views before softmax, whereas probability mean averages their normalized distributions. Hidden mean averages their last-layer representations before the shared linear LM head and therefore produces the same fused logits as logit mean; it serves as an implementation-level repeat. Best view selects the lower-entropy prediction separately at each token. Agreement weighting retains the logit-mean target but gives less weight to token positions where the averaged per-view distribution has high entropy. The first four variants fall within a narrow range of 27.2–29.2 pass@12, while agreement-weighted KL reaches 18.0.

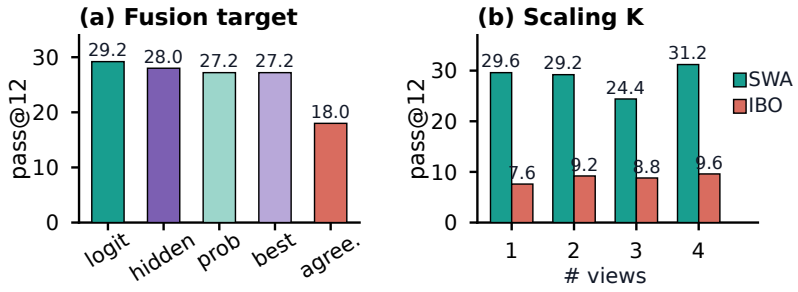


Figure 4: **Fusion target and view count. (a)** Fusion-level ablation on Qwen3-1.7B SWA with $K=2$ EN+ZH; logit averaging, hidden-state averaging, probability averaging, and best-view routing fall within 27.2–29.2 pass@12, while agreement-weighted KL reaches 18.0. **(b)** Scaling from EN to EN+ZH+ES+FR on Qwen3-1.7B SWA and IBO.

Nested language sets. We study teacher-view sets EN, EN+ZH, EN+ZH+ES, and EN+ZH+ES+FR. The four-view set achieves the best result on both SWA and IBO (Fig. 4b). We next analyze language-pair identity separately from this nested sequence.

5 Analysis

5.1 Can CMOPD transfer the benefit of multilingual test-time reasoning?

The central question of CMOPD is whether the benefit of multilingual test-time reasoning can be transferred to a single target-language policy. To test this, we compare how much adding a Chinese view improves an English-only baseline at inference time (English CoT versus EN+ZH synthesis) and during training (COPSD versus CMOPD).

Setting	Method	pass@12	Δ over EN
Training	COPSD (EN)	74.4	–
Training	CMOPD (EN+ZH)	79.2	+4.8
Inference	English CoT	81.2	–
Inference	Chinese CoT	77.2	–
Inference	EN+ZH synthesis	**85.6**	+4.4

Table 3: **Gain from adding a Chinese view during training or inference.** Results use Qwen3-8B on AfriMGSM-SWA. For EN+ZH synthesis, each trial generates independent English and Chinese reasoning traces and uses a third call to produce one Swahili answer. We compute pass@12 over 12 synthesized answers; the 24 source traces are not pooled.

EN+ZH synthesis uses independent English and Chinese reasoning traces followed by a synthesis call, improving pass@12 from 81.2 to 85.6. In the training comparison, CMOPD increases pass@12 over English-only COPSD from 74.4 to 79.2.

5.2 Where do equivalent views disagree?

To isolate the effect of language conditioning, we hold the problem and reference solution fixed and vary only the language of the teacher context. We fix a student rollout, query the same Qwen3-8B-Instruct teacher under English, Chinese, Spanish, and French contexts, and compute pairwise symmetric Kullback–Leibler divergence between the next-token distributions at every position. We then group the 15,360 Swahili rollout positions by token role.

Figure 3 shows that numbers, mathematical operators, and logical connectives have near-zero median divergence, whereas punctuation, LaTeX/Markdown, step markers, and meta phrases disagree more often. The all-view zero-divergence rates provide the same contrast: number and logical-connective positions match exactly at 20.1% and 16.6% of positions, compared with 0.6% for LaTeX/Markdown and 0.8% for punctuation. Thus, the views differ primarily in how the same reasoning step is expressed, rather than uniformly changing its mathematical content.

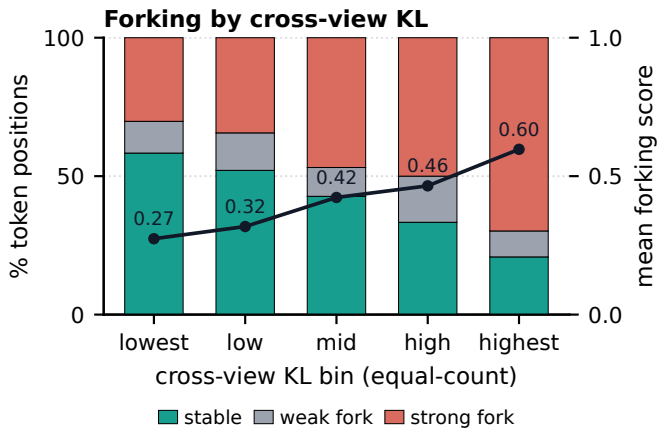


Figure 5: **Cross-view KL identifies unstable reasoning states.** We resample $M{=}12$ continuations at 480 token positions from 20 Qwen3-8B-Instruct SWA rollouts. Equal-count KL bins show that high-disagreement positions contain fewer prefixes whose continuations share the same correctness outcome; the black curve reports the mean forking score in each bin.

5.3 Does cross-view disagreement mark unstable reasoning states?

At each sampled token position, we test whether different continuations from the same prefix can change the final correctness outcome. We sample $M{=}12$ continuations from each prefix. A position is stable if all continuations have the same outcome, and unstable if correct and incorrect outcomes coexist. We quantify this variation as

$$\mathrm{fork}(t) = 4\hat{p}_t(1-\hat{p}_t), \quad \hat{p}_t = \frac{\#\text{correct continuations}}{12}.$$

We evaluate this score at 480 positions sampled from 20 Swahili rollouts.

Positions with higher cross-view disagreement have higher forking scores (Spearman $\rho{=}0.289$, $p{=}1.1{\times}10^{-10}$), and the association remains positive after controlling for relative position in the rollout ($\rho{=}0.171$, $p{=}1.7{\times}10^{-4}$). The within-category direction is also consistent for alphabetic, punctuation, and numeric tokens. Teacher views therefore disagree more at prefixes where the continuation is more likely to change the final correctness outcome.

Fusion normalization diagnostic. For geometric fusion $q(v) = \prod_k p_k(v)^{1/K}/Z$, the difference $\log q(y) - K^{-1}\sum_k \log p_k(y) = -\log Z$ is independent of token y. We therefore treat it only as a distributional-agreement diagnostic, not as evidence of beneficial token correction. Full results are in the supplementary material.

5.4 Does CMOPD preserve general capabilities?

To test whether mathematical OPD harms general instruction following, we evaluate the same Qwen3-8B Swahili checkpoints on IFEval (Zhou et al., 2023) with deterministic decoding. IFEval is not used for checkpoint selection.

Method	SWA	P-strict	P-loose	I-strict	I-loose	Δ_{IF}
Base 8B	69.20	81.7	85.2	87.8	90.0	0.0
COPSD $K{=}1$	74.4	**82.8**	**86.1**	**88.5**	**90.6**	-1.1
CMOPD $K{=}2$	**79.2**	82.4	85.8	88.0	90.3	-0.7

Table 4: **Instruction-following retention on Qwen3-8B Swahili checkpoints.** P/I denote prompt-/instruction-level accuracy; negative Δ_{IF} indicates improvement over Base.

CMOPD improves Swahili pass@12 over single-view COPSD while retaining general instruction-following ability: both trained checkpoints remain slightly above the base model on all four IFEval metrics. Thus, the task gain is not accompanied by measurable forgetting on this diagnostic.

5.5 What is the training overhead?

We compare the training cost of 100-step Qwen3-8B Swahili runs while varying only the teacher views. Table 5 shows that EN+ZH adds 2.8% over the single-view baseline, while four views add 7.4%. These runs show modest training overhead from the additional no-gradient teacher forwards; deployment remains unchanged.

Config.	Relative cost
$K{=}1$ (EN)	$1.000\times$
$K{=}2$ (EN, ZH)	$1.028\times$
$K{=}2$ (ES, FR)	$1.021\times$
$K{=}4$ (EN, ZH, ES, FR)	$1.074\times$

Table 5: **Relative training cost on Qwen3-8B Swahili.** All configurations use the same 100-step objective.

The supplementary material provides the qualitative example, complete token-category and target-shift results.

6 Limitations

We evaluate only on AfriMGSM, so whether the gains extend to non-mathematical reasoning remains unknown. All translated training views are generated by Qwen3-4B, and the results may depend on translation quality and model family. The main comparisons use English and Chinese views; larger view sets and alternative language pairs are examined only in limited settings. The test-time reasoning and token-level analyses focus on Qwen3-8B Swahili and may not generalize across models and target languages.

7 Conclusion

We presented CMOPD, which combines multiple language-conditioned predictions on a shared on-policy trajectory and distills them into a target-language policy. The trained model answers directly from the original low-resource query without multilingual decoding or answer synthesis. Across four AfriMGSM languages and three Qwen3 scales, EN+ZH CMOPD improves mean pass@12 over English-only COPSD at every scale, with heterogeneous effects across languages. On Qwen3-8B Swahili, adding the Chinese view improves training-time distillation by 4.8 points, while multilingual test-time synthesis improves over English-only reasoning by 4.4 points. Token-level analyses connect cross-view disagreement with unstable reasoning states and characterize how language-pair choice shapes the fused teacher distribution.

Ethics Statement

CMOPD aims to narrow the multilingual reasoning gap between high- and low-resource languages, with a current focus on four African languages (Swahili, Amharic, Yoruba, Igbo). Three points warrant disclosure. (i) **Translation provenance.** The Chinese, Spanish, and French "views" of each training example are produced by Qwen3-4B rather than by professional translators; residual translation errors could propagate into the distilled student and may interact differently with each target language. (ii) **Evaluation provenance.** The AfriMGSM test split we use is the publicly released human-translated split distributed by Adelani et al. (2024); we did not modify its annotations and we use it strictly for evaluation. (iii) **Evaluation scope.** The main table covers four target languages, with heterogeneous per-language effects (§4.2); the results do not constitute deployment evaluation for their language communities. No human subjects were involved; compute is reported in the supplementary material and training cost in §5.5.

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka

Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, 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 *arXiv preprint arXiv:2406.03368*, 2024.

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. xCoT: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*, 2024.

Aakanksha Chowdhery et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025. URL https://arxiv.org/abs/2501.17161.

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. *arXiv preprint arXiv:2509.25543*, 2025.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne 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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Berkcan Kapusuzoglu, Supriyo Chakraborty, Zain Sarwar, Chia-Hsuan Lee, and Sambit Sahu. Critique-guided distillation for robust reasoning via refinement. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2505.11628.

Yihong Liu, Raoyuan Zhao, Michael A. Hedderich, and Hinrich Schütze. Crosslingual on-policy self-distillation for multilingual reasoning. *arXiv preprint arXiv:2605.09548*, 2026.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Benjamin Muller, Deepanshu Gupta, Siddharth Patwardhan, Jean-Philippe Fauconnier, David Vandyke, and Sachin Agarwal. Languages you know influence those you learn: Impact of language characteristics on multi-lingual text-to-text transfer. *arXiv preprint arXiv:2212.01757*, 2022.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 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 (EMNLP)*, 2023.

Hugo Touvron et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Bibek Upadhayay and Vahid Behzadan. TaCo: Enhancing cross-lingual transfer for low-resource languages in LLMs through translation-assisted chain-of-thought processes. *arXiv preprint arXiv:2311.10797*, 2023.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin 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. *arXiv preprint arXiv:2402.07827*, 2024.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. MAD-OPD: Breaking the ceiling in on-policy distillation via multi-agent debate. *arXiv preprint arXiv:2605.01347*, 2026.

Xiaomi LLM-Core Team. MiMo-V2-Flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, Renjie Pi, Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nemotron-Cascade 2: Post-training LLMs with cascade RL and multi-domain on-policy distillation. *arXiv preprint arXiv:2603.19220*, 2026.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Zhiwei Yu, Tuo Li, Changhong Wang, Hui Chen, and Lang Zhou. Cross-lingual consistency: A novel inference framework for advancing reasoning in large language models. *arXiv preprint arXiv:2504.01857*, 2025.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. LinguaLIFT: An effective two-stage instruction tuning framework for low-resource language reasoning. *IEEE Transactions on Audio, Speech and Language Processing*, 34:1578–1593, 2026. doi: 10.1109/TASLPRO.2026.3671636.

Weihua Zheng, Xin Huang, Zhengyuan Liu, Tarun Kumar Vangani, Bowei Zou, Xiyan Tao, Yuhao Wu, Ai Ti Aw, Nancy F. Chen, and Roy Ka-Wei Lee. AdaMCoT: Rethinking cross-lingual factual reasoning through adaptive multilingual chain-of-thought. *arXiv preprint arXiv:2501.16154*, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv:2311.07911*, 2023.