Crosslingual On-Policy Self-Distillation for Multilingual Reasoning

Yihong Liu[1,2,*], Raoyuan Zhao[1,2,*], Michael A. Hedderich[1,2], Hinrich Schütze[1,2]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning (MCML)
[*]Equal contribution

June 2026

Outline

Imbalanced Multilingual Reasoning

- Large language models have made strong progress on **mathematical reasoning**.
- A key driver is the ability to generate **step-by-step reasoning traces**.
- However, this ability is not equally accessible across languages.
- Low-resource languages often receive:
  - less pretraining exposure,
  - less post-training supervision,
  - fewer high-quality reasoning examples.

Core Problem

A model may know how to solve a problem, but fail to access that ability when the problem is written, or a reasoning trace is generated in a low-resource language.

Multilingual Reasoning Is Imbalanced

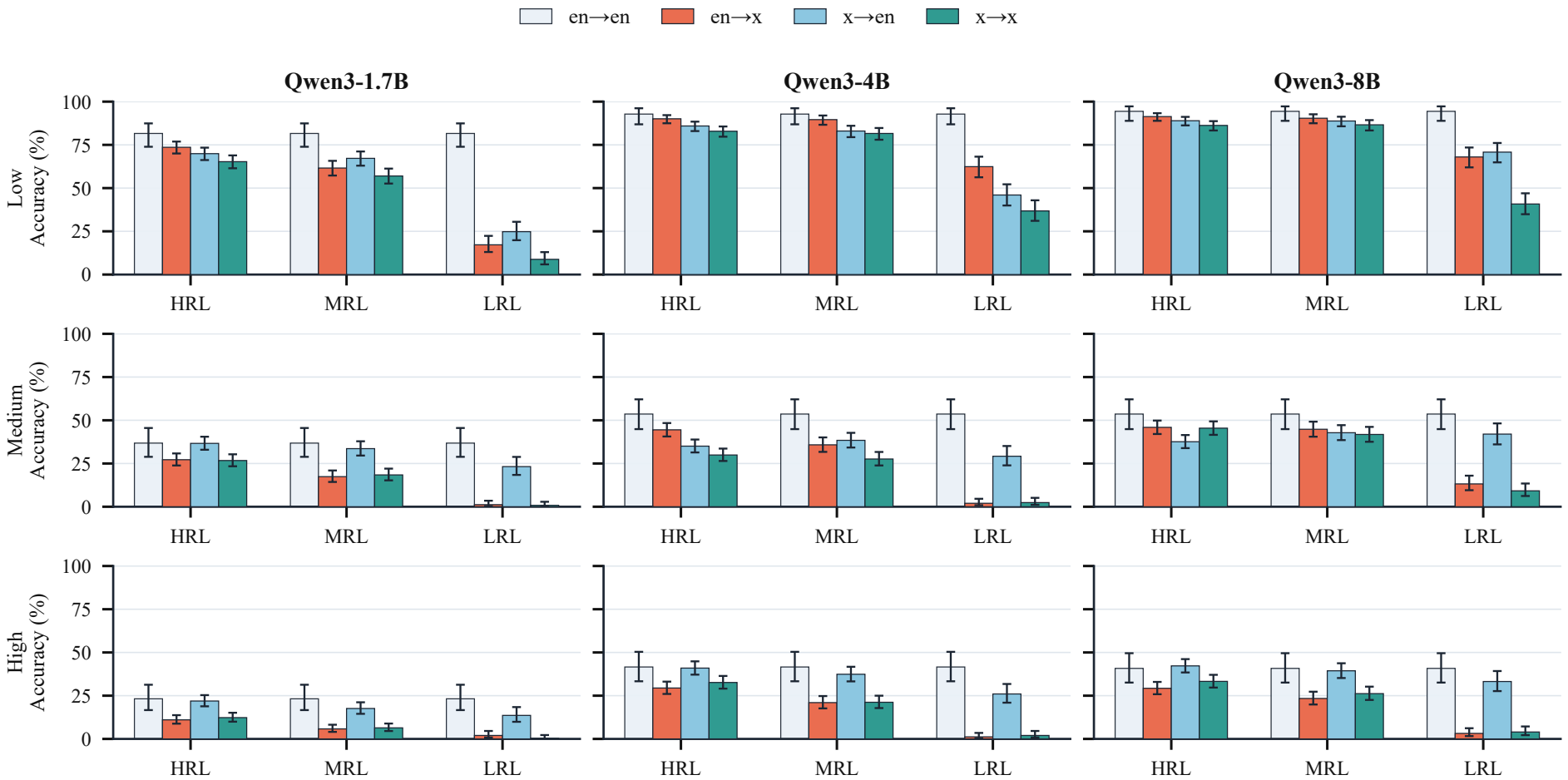


Figure: From Zhang et al. (2026a).

Imbalanced Multilingual Reasoning

High-resource setting

- Problem is written in English.
- Model has seen many similar examples.
- Reasoning traces are fluent and structured.
- Final answer is often formatted correctly.

Low-resource setting

- Problem is written in an underrepresented language.
- Model may understand parts of the input.
- Reasoning can become unstable, mixed, or repetitive.
- Final answer format may fail.

Question

How can we help the model express its latent reasoning ability in low-resource languages?

A Concrete Low-Resource Reasoning Failure

- Example: a Swahili reasoning trace for a simple problem.
- The correct computation should first apply the discount:

$$80 \times 0.8 = 64$$

and then add tax:

$$64 \times 1.1 = 70.4$$

- But the model:
  - misses the after-discount price,
  - applies tax to original price,
  - propagates the wrong state,
  - repeats the wrong answer,
  - fails to produce a valid final answer.

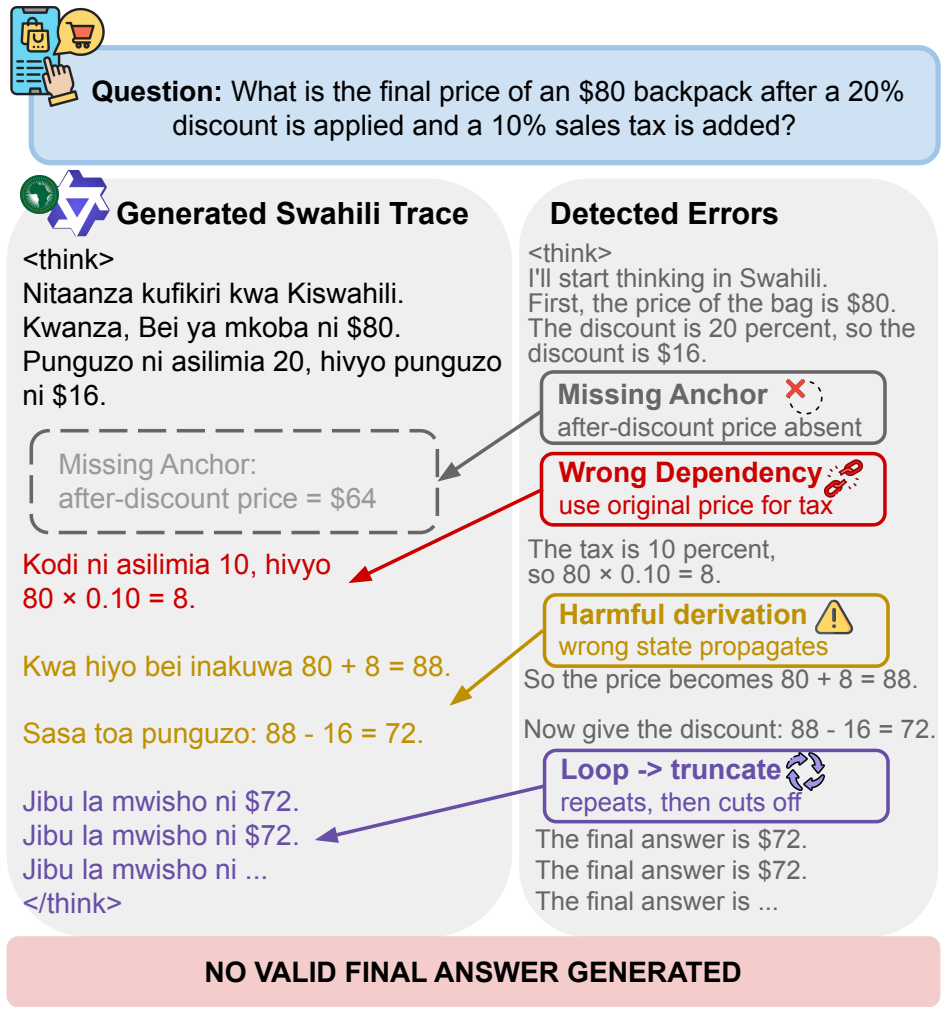


Figure: From Zhang et al. (2026a).

How to Improve Multilingual Reasoning?

- Existing approaches often try to transfer reasoning ability from high-resource languages.
- Common strategies include:
  - **Translate-and-test**: translate the input to English and solve it.
  - **Supervised fine-tuning**: train on translated reasoning traces.
  - **Reinforcement learning**: reward correct final answers.
  - **On-Policy Distillation**: use a stronger teacher to guide the model.

Our Direction

Use the model's **own English-accessible reasoning behavior** as supervision for low-resource reasoning.

Outline

Approach 1: Supervised Fine-Tuning

- Supervised fine-tuning trains the model to imitate reference outputs (Zhao et al., 2024; Zhang et al., 2024; Wu et al., 2025).
- For multilingual reasoning, a common recipe is:
  - start from English math problems,
  - translate questions into target languages,
  - translate or generate reasoning traces,
  - fine-tune the model on these traces.

Strength

SFT provides direct target-language demonstrations.

Limitation

It relies on fixed sampled traces, which may contain translation noise and may not match the model's own reasoning behavior at inference time.

Approach 2: Reinforcement Learning

- Reinforcement learning improves the model using rewards (She et al., 2024; Huang et al., 2025; Zhang et al., 2026c).
- For math reasoning, the reward is often outcome-based:
  - sample a response,
  - extract the final answer,
  - compare it with the gold answer,
  - reward correct answers.

Strength

RL does not require gold reasoning traces.

Limitation

In low-resource languages, correct answers may be rare, so binary rewards become sparse and provide little guidance for intermediate reasoning.

A Desirable Training Signal

- For low-resource reasoning, we want supervision that is:
  - **Dense**: feedback should be available at many reasoning steps.
  - **On-policy**: training should happen on the model's own generations.
  - **Scalable**: no need for expensive target-language rationales.

Key Idea

Instead of asking "Was the final answer correct?", ask: **At each step, how would the same model behave if it had useful English context?**

Preliminary: On-Policy Distillation

- On-policy distillation (OPD) trains a student on its **own generated trajectories** (Gu et al., 2024; Agarwal et al., 2024; Lu and Lab, 2025).
- A **stronger** model as a teacher provides token-level distributional feedback along those trajectories.
- This differs from **standard SFT**:
  - SFT imitates fixed reference traces.
  - OPD supervises the model on what it actually generates.
- This also differs from **outcome-based RL**:
  - RL often gives only a final reward.
  - OPD gives dense token-level feedback.

When Does On-Policy Distillation Work?

- OPD is not guaranteed to work simply because the teacher is stronger.
- Two **key conditions** for successful OPD (Li et al., 2026):
  - **Compatible thinking patterns**
    The student and teacher should reason in sufficiently **similar** ways. Large distributional mismatch can make token-level feedback hard to learn.
  - **Genuinely new capabilities**
    The teacher should provide **useful knowledge beyond** what the student already learned during training (pretraining, mid-training, post-training).

Implication

For OPD, choosing the teacher is not an easy job. The teacher must be both **compatible** with the student and **informative**.

Preliminary: On-Policy Self-Distillation

- On-Policy Self-Distillation addresses the compatibility issue by using the **same model** as both student and teacher.
- The two policies differ only in their conditioning context:
  - the **student** sees the normal inference-time prompt;
  - the **teacher** sees an enriched or privileged prompt.
- This keeps the teacher close to the student distribution while allowing privileged information to create a stronger learning signal.

Key Idea

Instead of finding a separate teacher that is both compatible and stronger, we make the same model stronger by giving it **better context**.

OPSD Objective

Given a problem x and privileged information y^*:

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

The student samples an on-policy trajectory:

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

At each step, we compare teacher and student distributions:

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

Intuition

The student learns how its own next-token predictions should change when useful privileged information is available.

Comparing Distillation Paradigms

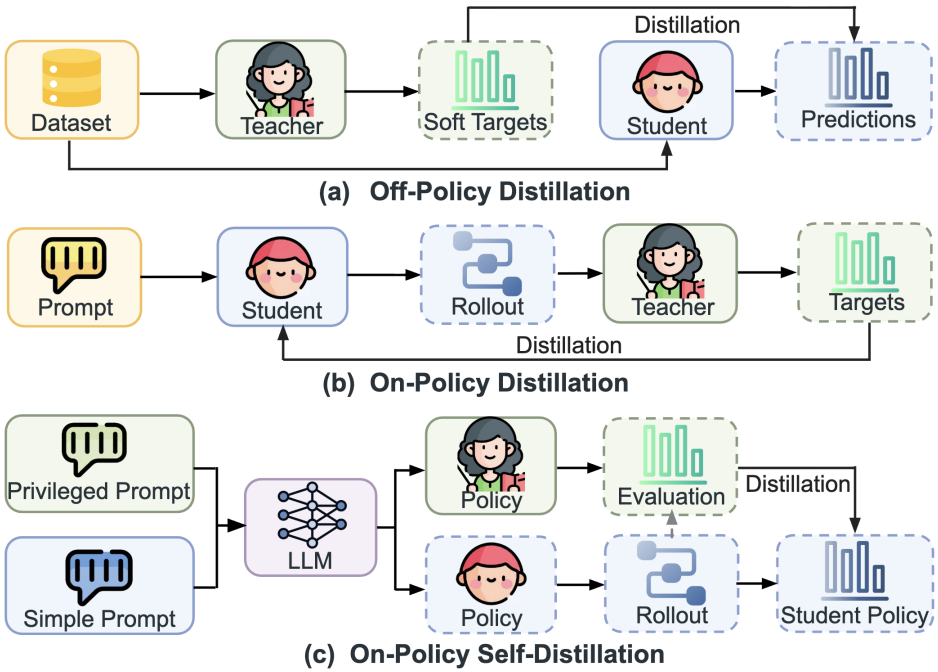


Figure: From Liu et al. (2026).

(a) Off-policy distillation

- Teacher provides targets from dataset examples.
- Student learns from **offline** supervision.
- Can suffer from **train–inference mismatch**.

(b) On-policy distillation

- Student first generates a **rollout**.
- Teacher evaluates the student's own trajectory.
- Better aligned with actual inference behavior.

(c) On-policy self-distillation

- Same model acts as both **teacher** and **student**.
- Different prompts provide different information.
- No external teacher is needed.

Why Extend OPSD to Multilingual Reasoning?

- OPSD has been used to improve reasoning by giving the teacher privileged information.
- Existing OPSD-style work mainly focuses on settings such as:
  - general mathematical reasoning (Zhao et al., 2026),
  - long-context reasoning (Zhang et al., 2026b),
  - compressed or efficient reasoning (Sang et al., 2026).
- But low-resource multilingual reasoning has a special challenge:
  - the model may reason well in English,
  - but fail to reason reliably in the target language.

Our Extension

Use English problem and solution as privileged context to improve low-resource reasoning.

Outline

Crosslingual On-Policy Self-Distillation

- We propose **COPSD**: *Crosslingual On-Policy Self-Distillation*.
- The student sees only the low-resource problem $x^{(L)}$:

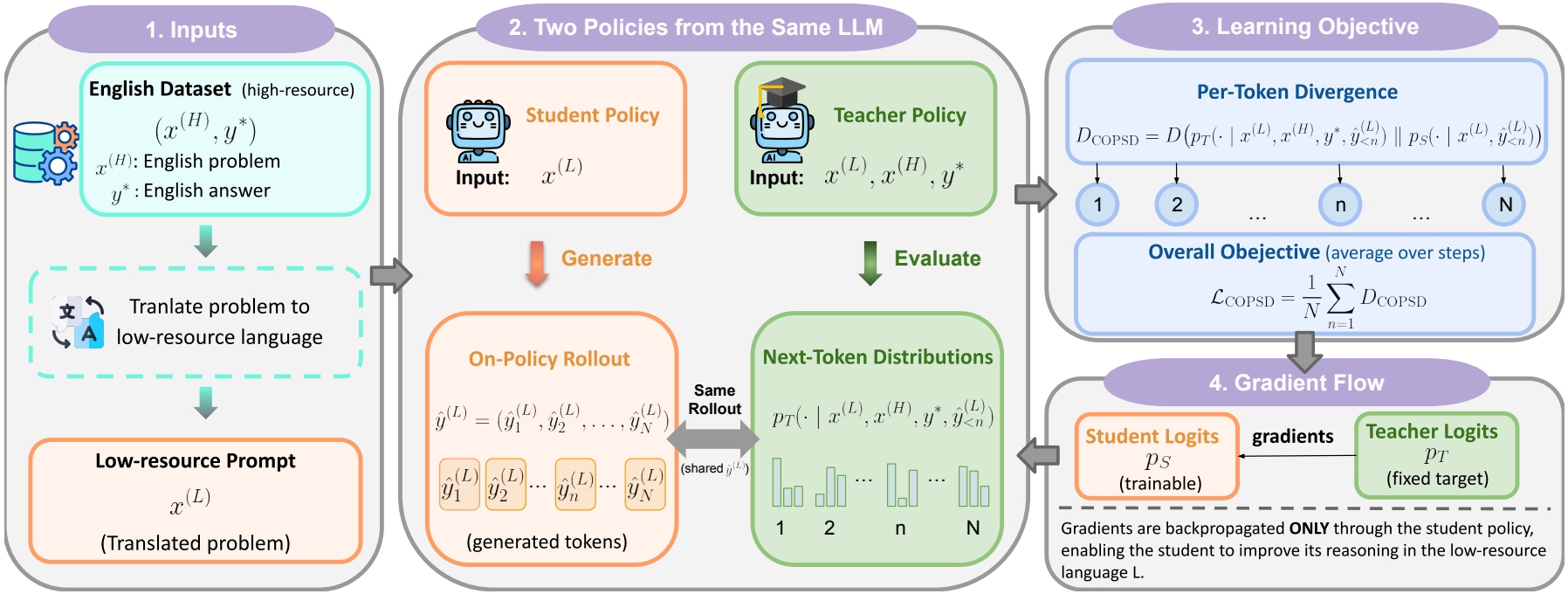
$$p_S(\cdot \mid x^{(L)}) = p_\theta(\cdot \mid x^{(L)}).$$

- The teacher sees privileged crosslingual information $x^{(H)}$ and y^*:

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

- Here:
  - $x^{(L)}$: low-resource problem,
  - $x^{(H)}$: English problem,
  - y^*: English reference solution.

COPSD Overview



- The student generates a **low-resource reasoning trajectory**.
- The teacher evaluates the same trajectory with **privileged context**.
- The model learns from **dense token-level distributional feedback**.

Why Does COPSD Help?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Compared with SFT

- Does not imitate one fixed sampled trace.
- Uses full teacher distributions.
- Trains on student-generated rollouts.

Compared with RL

- Does not rely only on final-answer rewards.
- Gives feedback at every decoding step.
- Avoids sparse low-resource reward signals.

Central Hypothesis

English context can make the same model behave like a stronger teacher for low-resource reasoning.

Training Data and Setup

Training Data

- Source: OpenThoughts math reasoning data.
- We sample 0.5K examples.
- Questions are translated into 17 African languages.
- English questions and solutions are used as privileged information.

Models

- `Qwen3-1.7B`
- `Qwen3-4B`
- `Qwen3-8B`

Important

COPSD does not require target-language reasoning rationales.

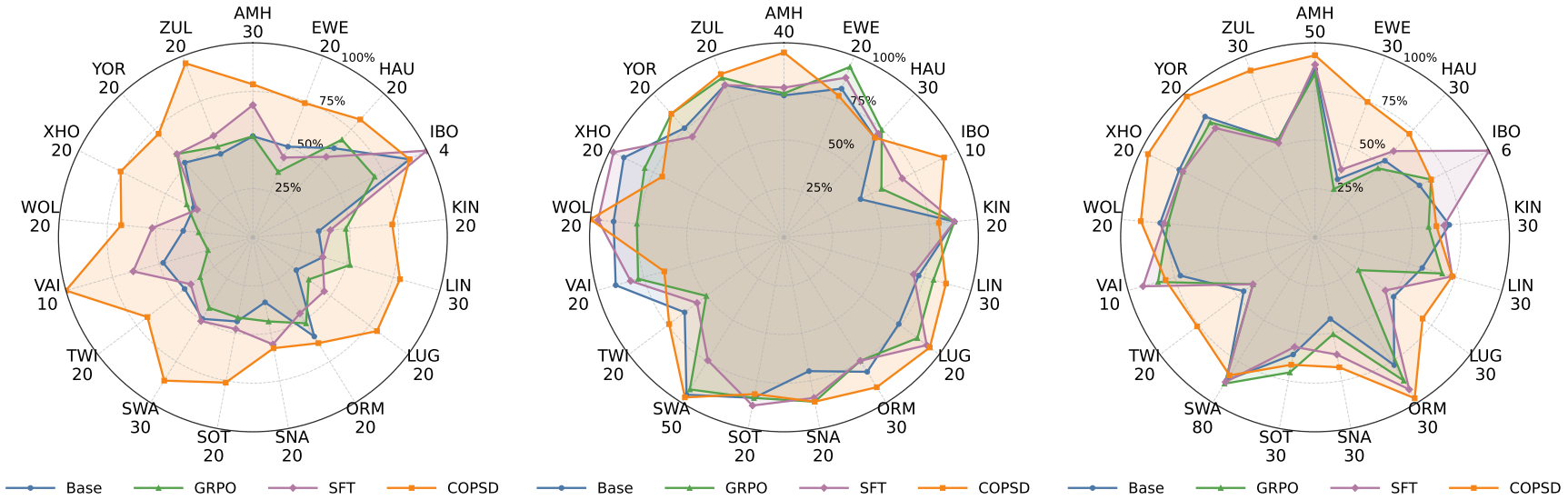
Evaluation

- Main benchmark: **AfriMGSM** (Adelani et al., 2025).
- It covers **17** low-resource African languages.
- Each language contains **250** math reasoning problems.
- Main metric: **Pass@12**.
  - For each problem, sample 12 responses and count the problem as solved if at least one response gives the correct final answer.
- Final answers are required in: `\boxed{}`.

Baselines

Base Qwen3, GRPO, and SFT (self-generated reasoning traces).

Main Result



- COPSD consistently improves over the base model.
- It also outperforms GRPO and SFT across most languages.
- Gains are especially strong for `Qwen3-1.7B` and `Qwen3-8B`.

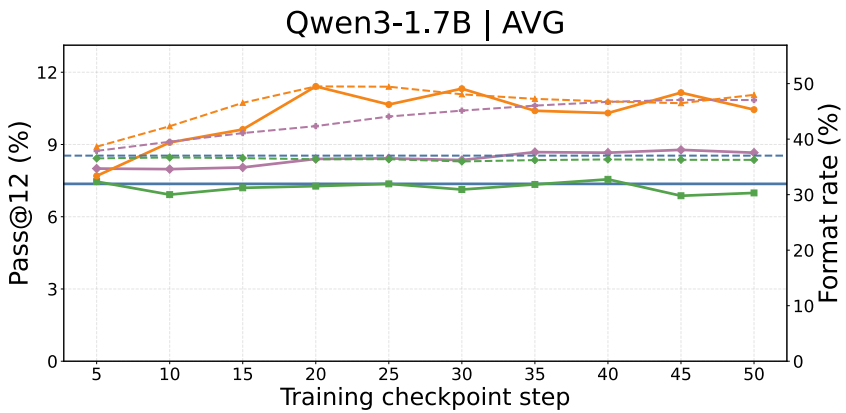
Average Pass@12 on AfriMGSM

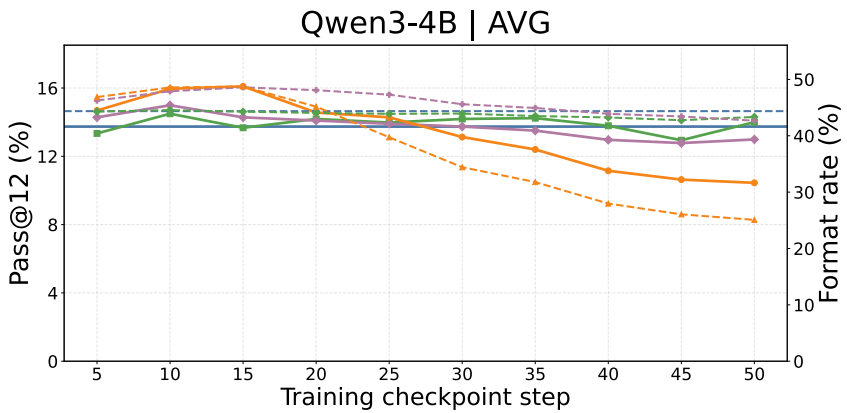
Model	Method	Avg. Pass@12
`Qwen3-1.7B`	Base	9.11
	GRPO	9.18
	SFT	10.12
	COPSD	**15.53**
`Qwen3-4B`	Base	19.20
	GRPO	19.36
	SFT	19.11
	COPSD	**20.61**
`Qwen3-8B`	Base	19.41
	GRPO	19.22
	SFT	20.42
	COPSD	**23.55**

Takeaway

Dense crosslingual self-distillation is more effective than sparse RL rewards or hard-token SFT targets.

Training Dynamics: COPSD Learns Quickly





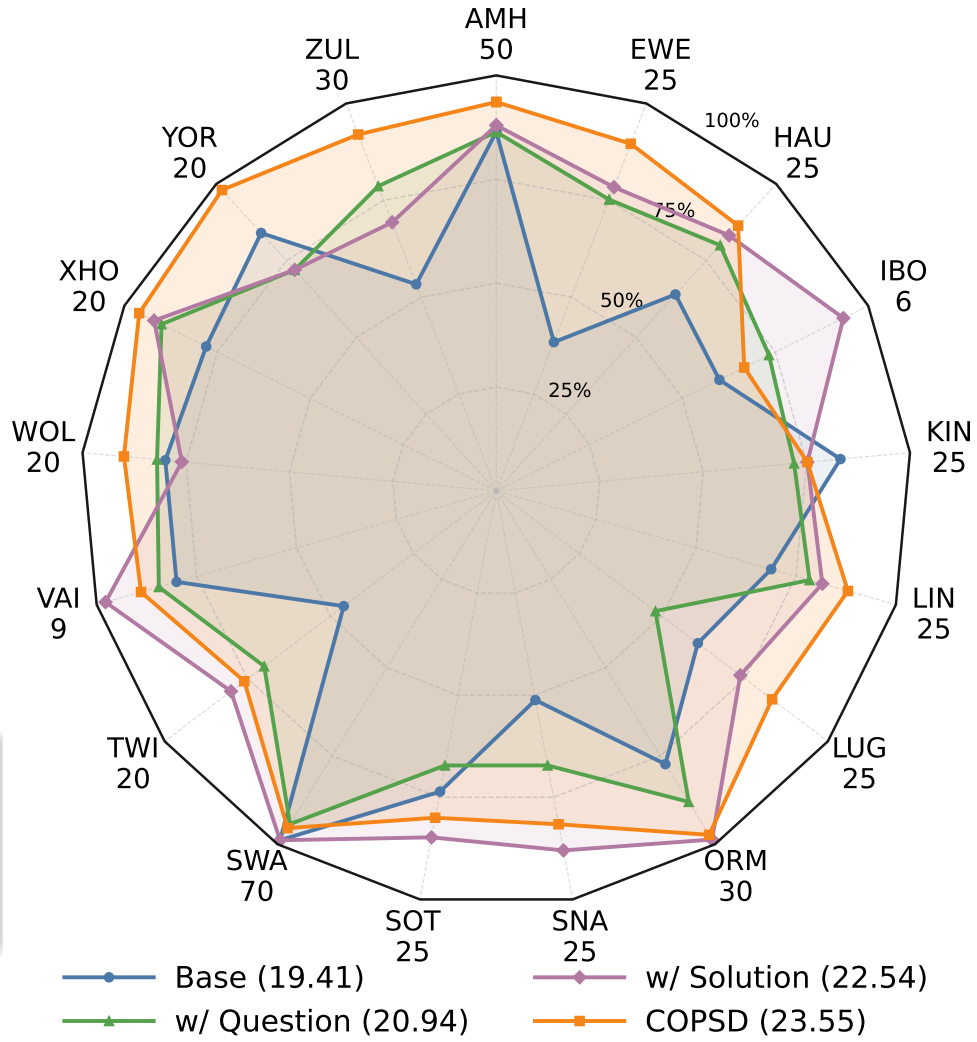
- COPSD improves Pass@12 and format rate in early training steps.
- GRPO and SFT show weaker or less stable improvement trends.
- Positive correlation between format rate and Pass@12 during training.

Ablation: What Privileged Information Matters?

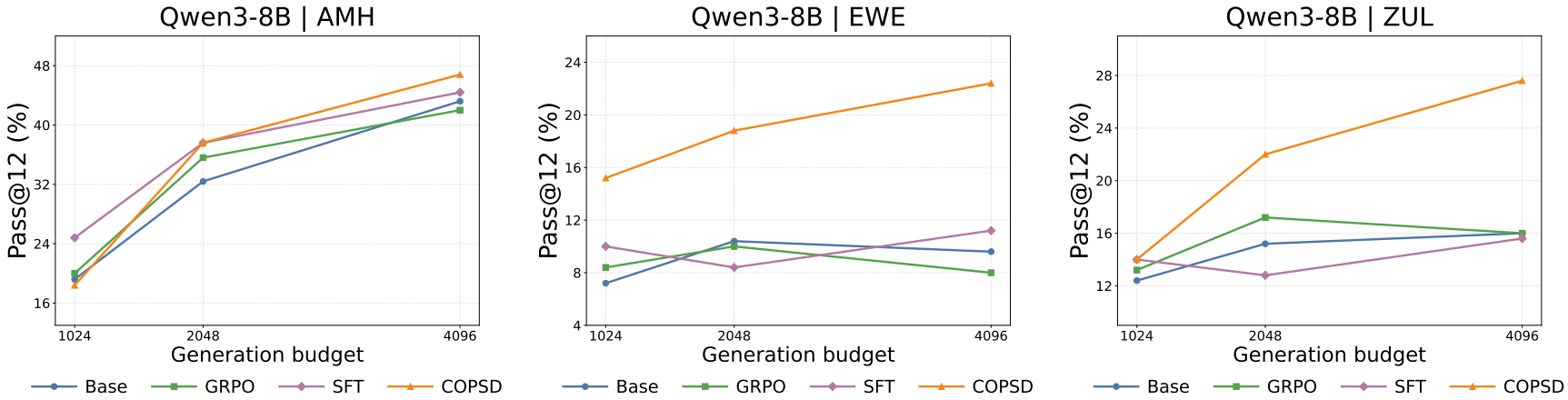
- The teacher receives:
  - English problem $x^{(H)}$,
  - English reference solution y^*.
- We compare:
  - full COPSD,
  - only English problem,
  - only reference solution.

Finding

Both sources help, but the full setting is strongest.



Test-Time Scaling

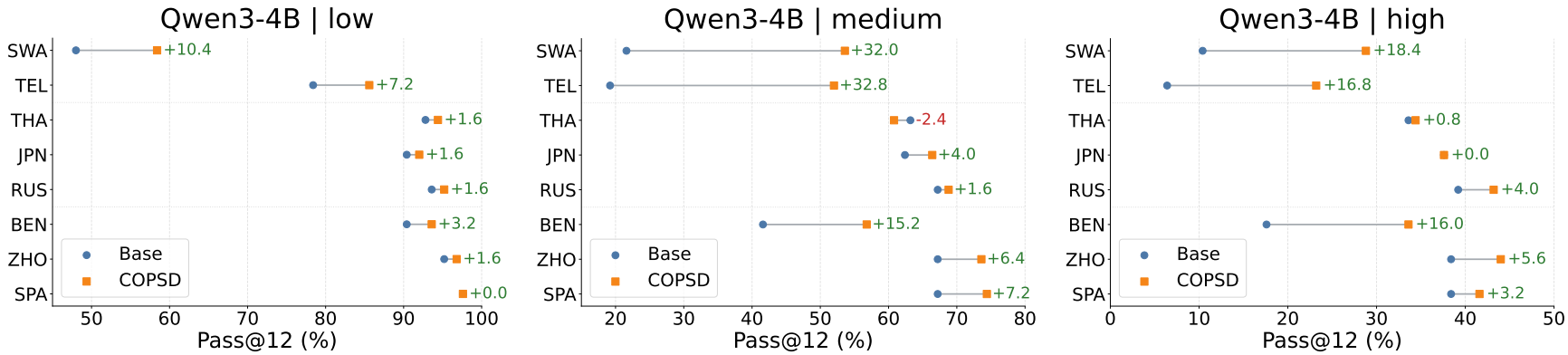


- Increasing the generation budget usually improves reasoning.
- COPSD benefits more clearly from larger budgets.
- COPSD strengthens the model's ability to use longer reasoning traces.

Example

For `Qwen3-8B`, COPSD improves from 18.12 to 23.55 average Pass@12 when the budget increases from 1,024 to 4,096 tokens.

Generalization to Harder Multilingual Reasoning



- We further evaluate on **PolyMath** (Wang et al., 2025), a harder multilingual reasoning benchmark with three difficulty levels.
- COPSD improves across low-, medium-, and high-difficulty settings.
- Gains are especially large for lower-resource languages.

Conclusion

(i) **Low-resource reasoning remains difficult.**
LLMs may possess reasoning ability but fail to access it in underrepresented languages.

(ii) **COPSD transfers English-accessible reasoning behavior.**
The student sees only the low-resource problem, while the teacher receives English privileged context.

(iii) **COPSD provides dense on-policy supervision.**
It trains on student-generated rollouts with token-level teacher distributions.

(iv) **COPSD improves multilingual mathematical reasoning.**
It outperforms Base, GRPO, and SFT on AfriMGSM and generalizes to PolyMath.

Takeaway

Crosslingual on-policy self-distillation is an effective path toward stronger multilingual reasoning, especially for low-resource languages.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Salomey Osei, Shamsuddeen Hassan Muhammad, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. 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.

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.

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

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.

Xinyu Liu, Darryl Cherian Jacob, Yang Zhou, Jindong Wang, and Pan He. 2026. Oisd: On-policy internal self-distillation of language models.

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

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. 2026. Crisp: Compressed reasoning via iterative self-policy distillation.

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.

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. 2025. Polymath: Evaluating mathematical reasoning in multilingual contexts.

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.

Jiaqiao Zhang, Zhoujun Li, Raoyuan Zhao, Jian Lan, Thomas Seidl, Michael A. Hedderich, Hinrich Schütze, and Yihong Liu. 2026a. Beyond

input understanding: Diagnosing multilingual mathematical reasoning with directed acyclic trace graphs.

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

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

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.

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