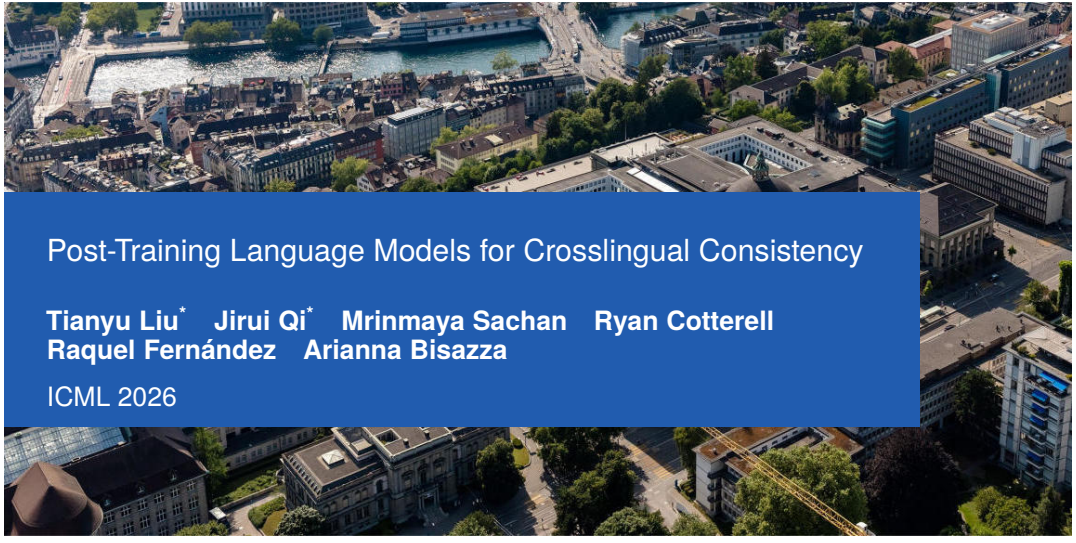
Rycolab, ETH Zürich
Univ. of Groningen & Univ. of Amsterdam



Post-Training Language Models for Crosslingual Consistency

Tianyu Liu* **Jirui Qi*** **Mrinmaya Sachan** **Ryan Cotterell**
Raquel Fernández **Arianna Bisazza**

ICML 2026

The same question, different answers

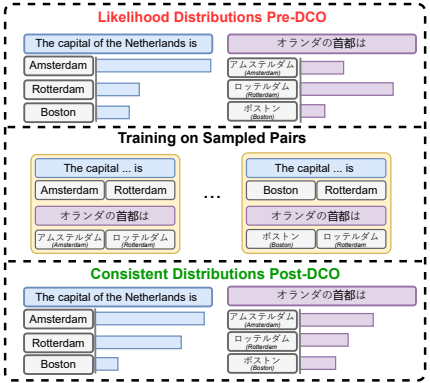
Same query, two languages → different answers

EN “What is the capital of Australia?” ⇒ Canberra ✔

SW “Mji mkuu wa Australia ni upi?” ⇒ Sydney ✖

- LLMs answer translation-equivalent prompts **inconsistently**.
- Undermines **reliability** and **equity** across languages — low-resource speakers get worse, unstable answers.

Goal. Post-train so the model answers the *same*, regardless of the prompt’s language.



Inconsistent rankings before alignment → aligned after.

Defining crosslingual consistency

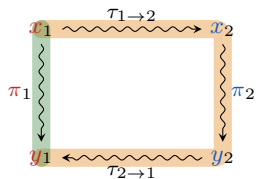
Two routes to a response distribution for a prompt in L_1:

- **Direct**: sample from $\pi_1(\cdot \mid x_1)$.
- **Round trip**: translate to L_2, respond, translate back: $\tau_{2\to1} \sharp \pi_2 \sharp \tau_{1\to2}$.

Consistent $\Leftrightarrow$ the two agree, within an f-divergence tolerance (up to a temperature T):

$$D_f\Big(\pi_1(\cdot \mid x_1) \,\Big\|\, (\tau_{2\to1} \sharp \pi_2 \sharp \tau_{1\to2})^T(\cdot \mid x_1)\Big) \leq \varepsilon$$

Zero divergence = the diagram commutes.



Direct vs **round trip** — both should land on the same y_1.

An objective: Penalized Consistency Optimization (PCO)

Per language, trade off consistency against drift from a fixed reference:

$$\mathcal{L}^{\text{PCO}}(\theta) = \underbrace{\mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big)}_{\text{fidelity}} \;-\; \underbrace{\beta \,\log\big(\tau_{2\to1} \sharp \pi_{\mathrm{ref}} \sharp \tau_{1\to2}\big)}_{\text{consistency reward}}$$

- **Closed-form optimum**: a weighted geometric mean

$$\pi^\star(\cdot \mid x_1) \;\propto\; \big(\tau_{2\to1} \sharp \pi_{\mathrm{ref}} \sharp \tau_{1\to2}\big)^{\beta_1}(\cdot \mid x_1) \;\cdot\; \pi_{\mathrm{ref}}(\cdot \mid x_1).$$

- With invertible translators and balanced strengths ($\beta_1\beta_2 = 1$), $\pi^\star$ is **exactly crosslingually consistent**.

From PCO to DCO: drop the roll-outs

Problem. The consistency term is an expectation under $\pi_\theta \Rightarrow$ optimizing PCO directly needs **on-policy roll-outs** (expensive, high-variance).

Idea (inspired by DPO). We *know* $\pi^\star$ in closed form — so just **regress the logits** onto it:

$$\mathcal{L}^{\text{DCO}}(\theta) = \left\| \ell_\theta(y_1 \mid x_1) - \beta_1 \log(\tau_{2\to1} \sharp \pi_{\text{ref}} \sharp \tau_{1\to2})(y_1 \mid x_1) - \log \pi_{\text{ref}}(y_1 \mid x_1) \right\|_1$$

- Same unique minimizer as PCO — but fully **off-policy**.
- Needs only *offline* evaluation on parallel data.
- Deterministic translators $\Rightarrow$ round trip = one eval of π_{ref} on translated input–output pairs. **No human labels.**

Experimental setup

Models. 9 LMs, 4 families, 3B–14B (Qwen2.5/3, Llama3, Gemma3, Aya-Expanse).

Data. 26 languages, 3 benchmarks:

- MMMLU — general knowledge
- XCSQA — commonsense reasoning
- BMLAMA — factual recall

Metrics.

- **CLC** (RankC): crosslingual consistency across language pairs.
- Accuracy on English and non-English.

Baselines. SFT, DPO (need labels); CALM (label-free).

Results: DCO improves consistency everywhere

MMMLU, joint N-lang	Qwen2.5-14B	Llama3.1-8B	Aya-Expanse-8B
Base (CLC)	68.6	60.9	72.2
+ CALM (label-free)	+4.2	+3.0	+1.4
+ DCO (label-free)	**+10.6**	**+9.4**	**+5.3**
+ DPO (labels)	+12.3	+10.1	+1.3
+ DPO→DCO (labels)	**+13.1**	**+13.8**	**+3.1**

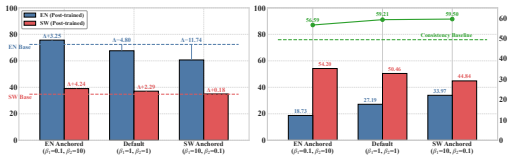
- Beats CALM; **matches/exceeds DPO without labels**; DPO→DCO best.
- Accuracy preserved: non-English ↑, English stable. Largest CLC gains on factual BMLAMA (+12 to +17).
- **Out-of-domain**: train 1 subject → ∼+11 CLC on unseen subjects.
- *Why accuracy rises*: geometric-mean ensemble down-weights uncertain, high-entropy (often wrong) experts.

Steering alignment toward low-resource languages

Strength parameters (β_1, β_2) with $\beta_1\beta_2 = 1$ are a **direction knob**.

- **English-anchored** (small β_{en}): keep English fixed, let the low-resource side move.
- English–Swahili: **+4.2** Swahili accuracy, English stable.
- Revises 54% of Swahili responses vs. only 19% of English.

Target low-resource languages *without* degrading the dominant one.



English–Swahili: accuracy (left), response change & CLC (right).

Takeaways

1. A precise, information-theoretic **definition** of crosslingual consistency (round-trip pushforward).
2. **PCO**: an objective whose closed-form optimum *is* consistent.
3. **DCO**: a cheap, **off-policy, label-free** surrogate with the same optimum — just regress the logits.

More consistent *and* more equitable multilingual models.

ETH zürich

Thank you!