ICML International Conference On Machine Learning

POSTER · 2026

PowerFlow

Unlocking the dual nature of LLMs by matching their own α-power distribution — label-free, one knob α

Ruishuo Chen · Yu Chen · Zhuoran Li · Longbo Huang✉

Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University · Beijing, China

清華大學 交叉信息研究院 Institute for Interdisciplinary Information Sciences, Tsinghua University

ff **Fortissimo — sharpen onto one solo line** `α = 4 · reasoning`

42.17

Top label-free reasoning · Qwen2.5-Math-7B avg@16 · above every RLIF baseline (TTRL 41.18) and matches supervised GRPO (42.38)

4.05

Strategy diversity on AIME24/25 · ahead of GRPO 3.93 & EMPO 3.80 — sharper, not narrower

α

ONE KNOB

the base model's own score
no external conductor · no labels

pp **Pianissimo — open into improvisation** `α = 0.5 · creativity`

0.33▲

Creativity restored · semantic diversity 0.23 → 0.33 (+45%) on Qwen2.5-1.5B-Instruct, α = 0.5

0.11▼

Less repetition · ROUGE-L redundancy 0.27 → 0.11 (lower is better) — while quality also rises 0.50 → 0.51

1 Why do reward-free methods keep collapsing?

Unsupervised RLIF elicits latent ability with no labels or verifier — but it leans on **heuristic intrinsic rewards** (self-certainty, majority voting) that have no well-defined optimization target.

- Existing RLIF (Intuitor, EMPO, TTRL) sharpens by *proxy* — with no target distribution to converge to.
- Under over-optimization the reward, not the model, sets the direction — so the elicitation inherits the reward's biases.
- The usual failure modes follow: length collapse or explosion, overconfidence, mode collapse.

Q: can one principled target elicit *both* reasoning and creativity, with no reward to hack?

`CODA` Match a distribution, not a reward — then there is nothing left to hack.

2 What can a model match with no labels?

PowerFlow recasts fine-tuning as **distribution matching** onto the base model's own **α-power (escort) distribution** — a self-derived reshaping, not a handcrafted reward. It modulates entropy while strictly preserving relative rankings, so fine-tuning stays grounded in pre-trained knowledge with no biased drift.

THE TARGET · ONE POWER KNOB

$$p_\alpha(y \mid q) \;\propto\; p_{\text{base}}(y \mid q)^\alpha$$

The map is monotone, so it re-weights mass while **preserving the base model's mode ranking**. A single knob then re-voices that instrument in two opposite directions:

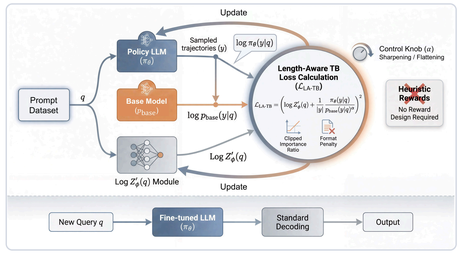
- `α > 1` sharpens onto a single solo line → reasoning.
- `α < 1` flattens it open → creative improvisation.

`CODA` α is the dynamic marking on the model's own score.

3 How do you match it without length collapse?

A GFlowNet is trained as an **amortized sampler** for the unnormalized target. But autoregressive log-probabilities shrink almost linearly with length, so naive matching is dominated by length rather than meaning — collapsing short (α > 1) or exploding into repetition (α < 1).

Our **length-aware Trajectory-Balance (LA-TB)** objective folds the partition function into a per-token energy, optimizing on a length-normalized surface so the gradient is driven by semantic density rather than sequence length.



Training matches the base model's α-power distribution via LA-TB; the α knob sets sharpen vs. flatten. Inference is ordinary decoding — no reward model.

`CODA` One knob, no external conductor — fully label-free.

4 Does the length fix distort the target?

No — length-awareness is principled, not a heuristic patch. LA-TB is the **I-projection**: among all length-calibrated distributions it is the *minimum-distortion* correction of the target, and its global KL distortion is **second-order** in the length multiplier λ_q.

Measured: ≈ 91% of α-power pairwise rankings are preserved in practice (inversion rate ≈ 0.09, trained Qwen2.5-Math-1.5B).

`CODA` Structure survives — the instrument is retuned, not replaced.

5 α > 1: does sharpening beat supervised RL?

avg@16 (Average over MATH500, Olympiad, AIME24/25, AMC23, GPQA). α = 4 for base models, α = 2 for instruct.

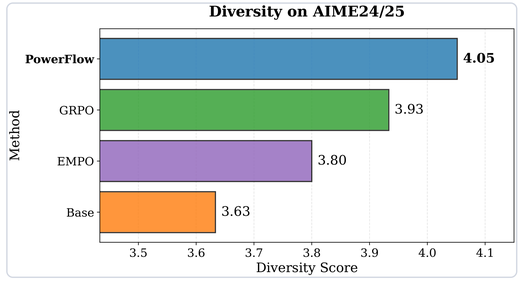
Model	Best RLIF	PowerFlow	GRPO$^{sup.}$
Qwen2.5-1.5B	18.95	19.85	18.13
Qwen2.5-Math-1.5B	32.45	34.30	32.75
Llama-3.2-3B-Inst	22.37	22.88	22.33
Qwen2.5-Math-7B	41.18	42.17	42.38

PowerFlow tops every label-free baseline at all four scales, and beats supervised GRPO on three — comparable on Math-7B (42.17 vs 42.38, within 1σ).

`CODA` Unsupervised, yet it rivals supervised RL.

6 Does concentrating the mass narrow the model?

No. The α-power map rescales the *whole* density and keeps the base model's multi-modal structure, so sharpened models retain diverse solution strategies.

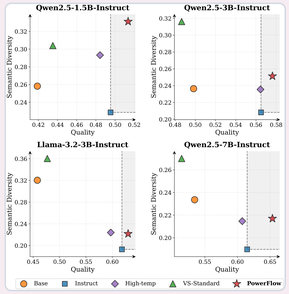


Distinct-strategy diversity on AIME24/25 (LLM-judged). PowerFlow 4.05 leads GRPO 3.93 and EMPO 3.80.

`CODA` Sharper, not narrower.

7 α < 1: can the same knob free creativity?

Alignment implicitly over-sharpens aligned models ("typicality bias"), suppressing creative modes. Flattening with `α = 0.5` counteracts it, reviving latent expressive paths.



Quality vs. semantic diversity, four instruct models. PowerFlow (★) sits in the Pareto-improvement region — the only method to lift both at once.

`CODA` Best of both — expressive, and still on-instruction.

8 Which ingredient does the work?

Component ablation on Qwen2.5-Math-1.5B, avg@16 across all six reasoning benchmarks.

Configuration	Avg
Distribution matching only (TB-traj)	≈ 0
+ format penalty only	14.30
+ length correction only	33.48
PowerFlow (full)	34.30

The length-aware surface supplies the bulk (33.48); the format penalty adds +0.82. At 7B a format-only variant reaches 36.22 — still 5.95 below PowerFlow's 42.17.

`CODA` The length-aware surface is what makes matching stable.