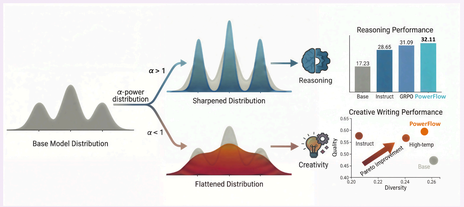
PowerFlow: Unlocking the Dual Nature of LLMs via Principled Distribution Matching

ICML International Conference On Machine Learning
2026 · POSTER

One tunable knob α — sharpen to reason, flatten to create.

Ruishuo Chen · Yu Chen · Zhuoran Li · Longbo Huang✉
Institute for Interdisciplinary Information Sciences · Tsinghua University · Beijing, China
Institute for Interdisciplinary Information Sciences, Tsinghua University

Code & Models



FRAMEWORK **PowerFlow** matches a length-aware α-power target of the base model. A single knob α directionally elicits **reasoning** ($\alpha>1$, sharpen) or **creativity** ($\alpha<1$, flatten) — fully unsupervised under standard decoding.

+1.7 avg@16 over GRPO (Qwen2.5-1.5B)

Stable monotonic gains; no length collapse

Pareto quality × diversity across 4 models

I-proj. minimum-KL length correction

1 Motivation

Unsupervised **RLIF** elicits latent LLM capabilities, but its *heuristic* rewards cause: **length collapse**, **overconfidence**, **mode collapse**, & reward hacking.

Q: Replace handcrafted rewards with a principled, controllable target distribution?

2 Core Idea: α -Power

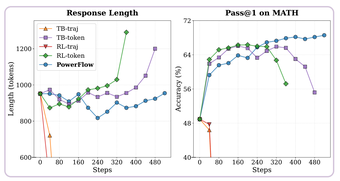
Recast fine-tuning as **distribution matching**. Target the α-power (escort) distribution of the base model — entropy modulates while **rankings & mode structure are strictly preserved**:

TARGET DISTRIBUTION

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

$\alpha > 1$	$\alpha = 1$	$\alpha < 1$
Sharpen · Reason	Base · Identity	Flatten · Create

3 Why Length-Aware? ★ Evidence



Naïve TB / RL: length collapse at $\alpha>1$.

Token TB: repetition decay at $\alpha<1$.

PowerFlow stays length-stable with monotonic gains on MATH.

4 GFlowNet as Amortized Sampler

Bypass the intractable $Z(q)$ via **Trajectory Balance** (TB), a variational surrogate that at equilibrium of Z_ϕ satisfies $\mathbb{E}[\nabla_\theta \mathcal{L}_{\text{TB}}] = 2\nabla_\theta \mathbb{D}_{\text{KL}}(P_F \| p_{\text{target}})$.

STANDARD TB (AUTOREGRESSIVE)

$$\mathcal{L}_{\text{TB}} = \left(\log Z_\phi + \sum_{t=1}^{T} \log \pi_\theta(y_t \mid y_{<t}, q) - \log \tilde{p}_{\text{target}}\right)^2$$

But: on LLMs this is **dominated by sequence length**, not semantic density — causing *length collapse* ($\alpha > 1$) or *repetition explosion* ($\alpha < 1$).

5 Length-Aware TB ★ KEY

Reparameterize $Z_\phi(q, y) = \left(Z'_\phi(q)\right)^{|y|}$ and normalize by $|y|$ — projecting onto a length-normalized energy surface:

POWERFLOW LOSS (w: CLIPPED IS; ψ: FORMAT PENALTY)

$$\mathcal{L}_{\text{PF}} = w\left(\log Z'_\phi + \frac{1}{|y|}\log \pi_\theta - \alpha\left[\frac{1}{|y|}\log p_{\text{base}} + \psi\right]\right)^2$$

The induced target is a 1-D exponential tilt of p_α by $|y|$: $\pi^\star \propto p_{\text{base}}^\alpha \, e^{-\lambda_q |y|}$.

Theory. LA-TB is the *I-projection* — the minimum-KL length correction onto length-calibrated distributions.

Empirical: pairwise inversion rate vs. the ideal α-power target on Qwen2.5-Math-1.5B is **IR ≈ 0.09** — ~91% of rankings preserved.

6 Reasoning ($\alpha > 1$) ★ GRPO-Matched

Avg@16 on MATH500 · Olympiad · AIME24/25 · AMC23 · GPQA

Method	Avg.	Δ GRPO
Qwen2.5-1.5B		
Intuitor	18.95	+0.82
PowerFlow	**19.85**	**+1.72**
GRPO (sup.)	18.13	ref
Qwen2.5-Math-1.5B		
EMPO	32.45	−0.30
PowerFlow	**34.30**	**+1.55**
GRPO (sup.)	32.75	ref
Llama-3.2-3B-Inst.		
PowerFlow	**22.88**	**+0.55**
GRPO (sup.)	22.33	ref
Qwen2.5-Math-7B		
EMPO / TTRL	40.88 / 41.18	−1.50 / −1.20
PowerFlow	**42.17**	**−0.21**
GRPO (sup.)	42.38	ref

7 Reasoning Diversity Preserved

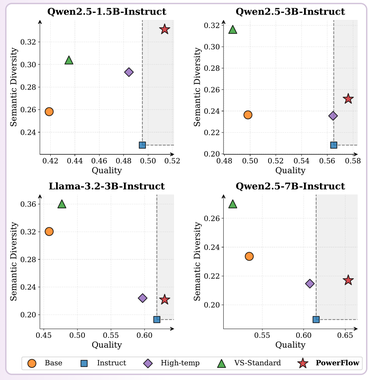
4.05 PowerFlow

3.93 GRPO (sup.)

3.80 EMPO (RLIF)

Strategy-diversity on AIME24/25 (DeepSeek-V3.2 judge, 16 samples/problem). The α-power transform preserves the base model's multi-modal structure — **no monolithic mode collapse**.

8 Creativity ($\alpha < 1$) ★ Pareto-Dominant

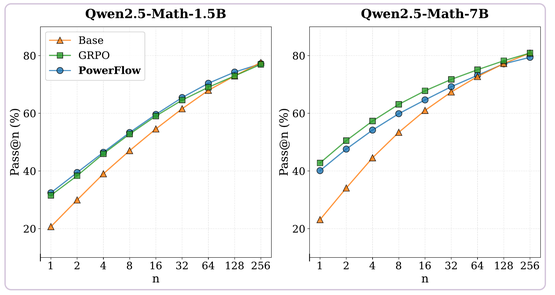


AXES
Quality × Diversity
↗ upper-right is better

★ RESULT
Beyond the Instruct frontier

SETUP
4 model families
$\alpha = 0.5$ flatten

9 Reasoning Mechanism: Pass@n



PowerFlow tracks GRPO across $n = 1 \to 256$ on **OlympiadBench** — sharpening lifts pass@1 *without* sacrificing reasoning coverage.

10 Takeaways

PRINCIPLED. One knob α replaces all heuristic rewards.

LENGTH-AWARE. I-projection neutralizes length bias.

DUAL-NATURE. Reasoning & creativity, one loss.

PRACTICAL. ~10% step overhead; standard decoding.

PowerFlow · ICML 2026 · Acknowledgements: NSFC Grant 52494974.

Code: **github.com/Chenruishuo/PowerFlow** · Contact: **crs25@mails.tsinghua.edu.cn**