ICML 2026 · UNSUPERVISED CAPABILITY ELICITATION

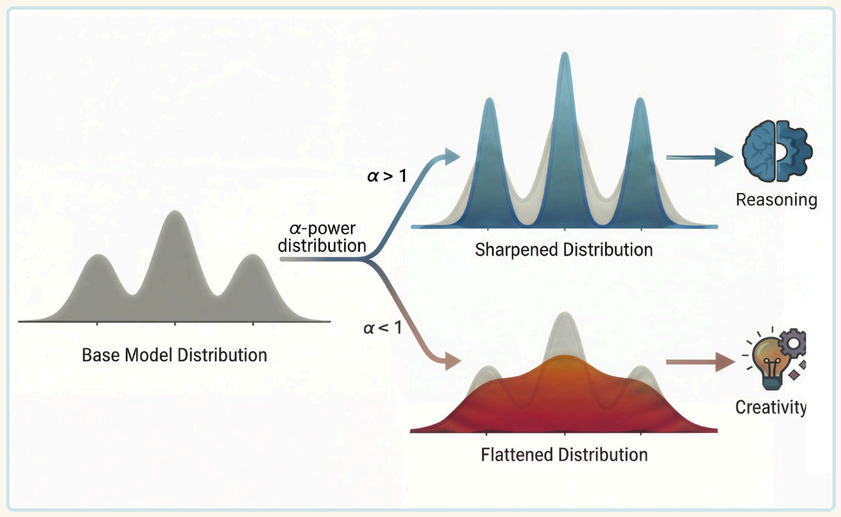
PowerFlow

Unlocking the Dual Nature of LLMs via Principled Distribution Matching

Ruishuo Chen · Yu Chen · Zhuoran Li · Longbo Huang✉
IIIS, Tsinghua University · ✉ longbohuang@tsinghua.edu.cn

I THE LANDSCAPE
One base model, two natures

A pretrained model is already a rugged landscape of output modes. **PowerFlow** reshapes it toward the **α-power (escort) distribution** $p_\alpha(y|q) \propto p_{\text{base}}(y|q)^\alpha$ — one exponent sets the direction: **sharpen** (α > 1) for reasoning, **flatten** (α < 1) for creativity. No external rewards, labels, or verifiers.



Sharpen raises the peaks (reasoning); **flatten** spreads mass into the basins (creativity) — reshaping elevation, preserving the tallest peaks.

42.17
PowerFlow's best avg@16 · Qwen2.5-Math-7B

+0.99
over TTRL, the strongest RLIF baseline

4.05
reasoning-path diversity (1–5), AIME24/25

3 / 4
tested models beat supervised GRPO

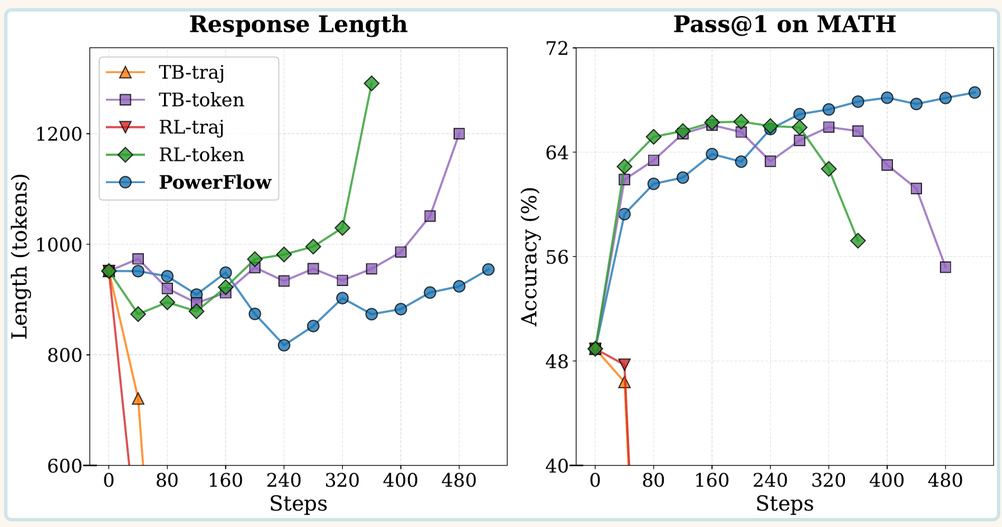
II LENGTH-AWARE DISTRIBUTION MATCHING
Match the α-power surface, not a reward

Heuristic RLIF rewards carry no principled target, so over-optimization drifts into **length and mode collapse**. PowerFlow instead matches a target distribution with a GFlowNet Trajectory-Balance objective; the escort target rescales the surface while **preserving the base model's mode rankings**.

Autoregressive log-probabilities shrink with length, so naive matching collapses or explodes it. PowerFlow reparameterizes the partition function as a length-aware energy term $Z_\phi(q,y) = (Z'_\phi(q))^{|y|}$ and optimizes on a **length-normalized surface**:

LENGTH-AWARE TRAJECTORY BALANCE

$$\mathcal{L}_{\text{LA-TB}} = \left(\log Z'_\phi(q) + \frac{1}{|y|} \log \frac{\pi_\theta(y|q)}{\tilde{p}_{\text{target}}(y|q)} \right)^2$$



Naive matching fails: -traj objectives collapse length, -token ones inflate repetition then decay. PowerFlow holds length stable and accuracy rising.

III SHARPEN · α = 4 (base), 2 (instruct) / FLATTEN · α = 0.5
The two natures, measured

RAISING THE RIDGES — REASONING

Across four models and six benchmarks (avg@16), PowerFlow tops the strongest RLIF baseline in every model block and beats **supervised GRPO on three of four** — PowerFlow comes out level on the 7B (42.17 vs 42.38).

Model	Top RLIF UNSUP.	PowerFlow OURS	GRPO SUPERV.
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**

Variety preserved. Sharpening the whole surface keeps solution diversity highest — **4.05** on AIME24/25, above GRPO (3.93) and EMPO (3.80).

FLOODING THE BASINS — CREATIVITY

Alignment over-sharpens instruct models (a *typicality bias*). Flattening recovers atypical, high-quality paths — simultaneously lifting **both diversity and quality** above Instruct on all four.

Instruct model	Diversity INS → OURS	Quality INS → OURS
Qwen2.5-1.5B	0.23 → **0.33**	0.50 → **0.51**
Qwen2.5-3B	0.21 → **0.25**	0.57 → **0.58**
Llama-3.2-3B	0.19 → **0.22**	0.62 → **0.63**
Qwen2.5-7B	0.19 → **0.22**	0.61 → **0.65**

Best of both worlds. Temperature or prompt tricks trade quality for variety; PowerFlow is the **only** evaluated method that lifts both (poem, joke, story avg).