POWERFLOW

Unlocking the Dual Nature of LLMs via Principled Distribution Matching

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

IIIS, Tsinghua University

Existing unsupervised RL (RLIF) leans on **hand-crafted intrinsic rewards** with no principled target. PowerFlow instead **matches the α-power (escort) distribution** — one exponent α *tempers* the model: **quench (α>1)** for reasoning, **heat (α<1)** for creativity.

42.17 AVG@16 · BEST RLIF

4.05 REASONING DIVERSITY · > GRPO

1 EXPONENT, TWO NATURES

THE CONTROL KNOB · ONE EXPONENT α

TEMPER THE DISTRIBUTION

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

The **α-power (escort) distribution** of statistical mechanics — it modulates entropy while **preserving relative mode rankings**.

α > 1 · Quench

Sharpen — concentrate mass onto latent **reasoning** paths, surfaced without new supervision.

→ REASONING

α = 1 · BASE MODEL

α < 1 · Heat

Flatten — release the variety that alignment over-sharpened away, recovering suppressed modes.

→ CREATIVITY

Reshapes the **whole trajectory density** — provably **not** the same as lowering the decoding temperature (the Low-temp baseline it beats).

01 FROM HEURISTIC REWARDS TO A PRINCIPLED TARGET

RLIF elicits latent ability **without labels or verifiers**, but hand-crafted rewards chase no principled target — drifting into length **collapse/explosion** and mode collapse.

PowerFlow matches the **α-power target**: an order-preserving reshaping that leaves the base model's rankings and modes intact — **grounded in pretraining**.

02 QUENCH → REASONING (α > 1)

Sharpening (α = 4) surfaces latent reasoning that temperature and instruction tuning miss — **best RLIF, level with supervised GRPO**.

QWEN2.5-MATH-7B	AVG@16
PowerFlow (ours)	**42.17**
TTRL — best prior RLIF	**41.18**
EMPO	**40.88**
GRPO — supervised ref	***42.38***

Six-benchmark avg@16; beats supervised GRPO on **3 of the 4 tested models**.

> SOLUTION DIVERSITY · AIME24/25 (1–5 SCALE)
>
> **4.05** PowerFlow › 3.93 GRPO › 3.80 EMPO — distinct-strategy score, DeepSeek-V3.2 judge.

03 LENGTH-AWARE TRAJECTORY BALANCE

Cast GFlowNet as an **amortized variational sampler** for the unnormalized α-power density. Autoregressive length bias otherwise dominates the gradient, so PowerFlow optimizes a length-normalized surface:

$$\pi^\star(y \mid q) \propto p_{\text{base}}(y \mid q)^{\alpha} \,/\, Z'_\phi(q)^{|y|}$$

The $Z'_\phi(q)^{|y|}$ term neutralizes length bias, giving **stable, overall-upward** gains where raw matching collapses. **α = 4** is the base-model sharpening sweet spot (sweep {2, 4, 6}).

04 HEAT → CREATIVITY (α < 1)

On aligned instruct models, flattening (α = 0.5) restores suppressed variety — a **Pareto-dominant shift**, and the **only evaluated method** to also improve generation quality.

Δ VS INSTRUCT	SEM. DIV.	QUALITY
Qwen2.5-1.5B	**+0.103**	**+0.018**
Qwen2.5-3B	**+0.043**	**+0.011**
Llama-3.2-3B	**+0.029**	**+0.012**
Qwen2.5-7B	**+0.027**	**+0.039**

Overall-average over poem/joke/story: every family rises on **both** semantic diversity and quality — a Pareto gain, never a quality-for-diversity trade.

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

Code: github.com/Chenruishuo/PowerFlow

GET THE CODE
scan for the repo