PowerFlow

Unlocking the Dual Nature of LLMs via Principled Distribution Matching

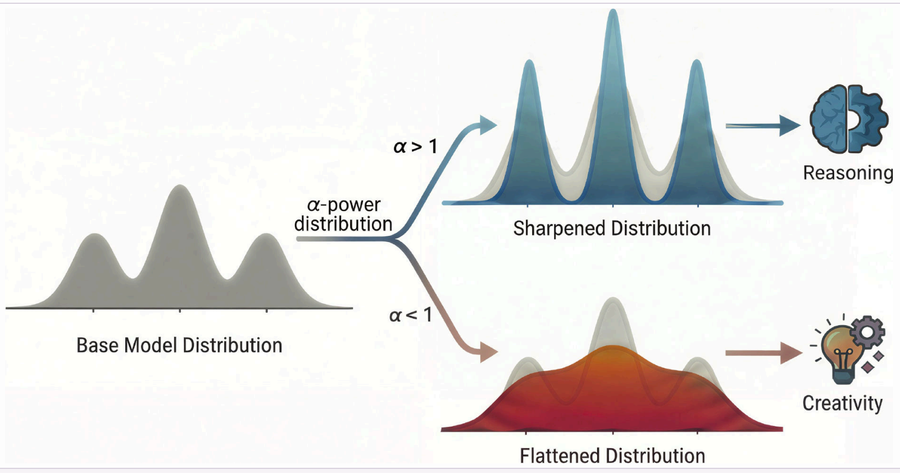
Ruishuo Chen Yu Chen Zhuoran Li Longbo Huang✉

IIIS, Tsinghua University · ✉ Corresponding author

ICML International Conference On Machine Learning

ONE DIAL · TWO NATURES

PowerFlow matches the α-power distribution of a model's own outputs, so one knob α elicits its dual nature — sharpen for reasoning, flatten for creativity.



0 < α < 1
Flatten → Creativity
only evaluated method to lift diversity *and* quality

α = 1
Base model

α > 1
Sharpen → Reasoning
42.17 avg — tops RLIF (**TTRL 41.18**)

01 Distribution matching, not reward design

RLIF elicits reasoning from intrinsic rewards — no labels — but those rewards are **handcrafted**, with no principled target, so over-optimization produces length collapse, overconfidence, and mode collapse.

PowerFlow instead matches the model's own α-power (escort) distribution:

TARGET DISTRIBUTION

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

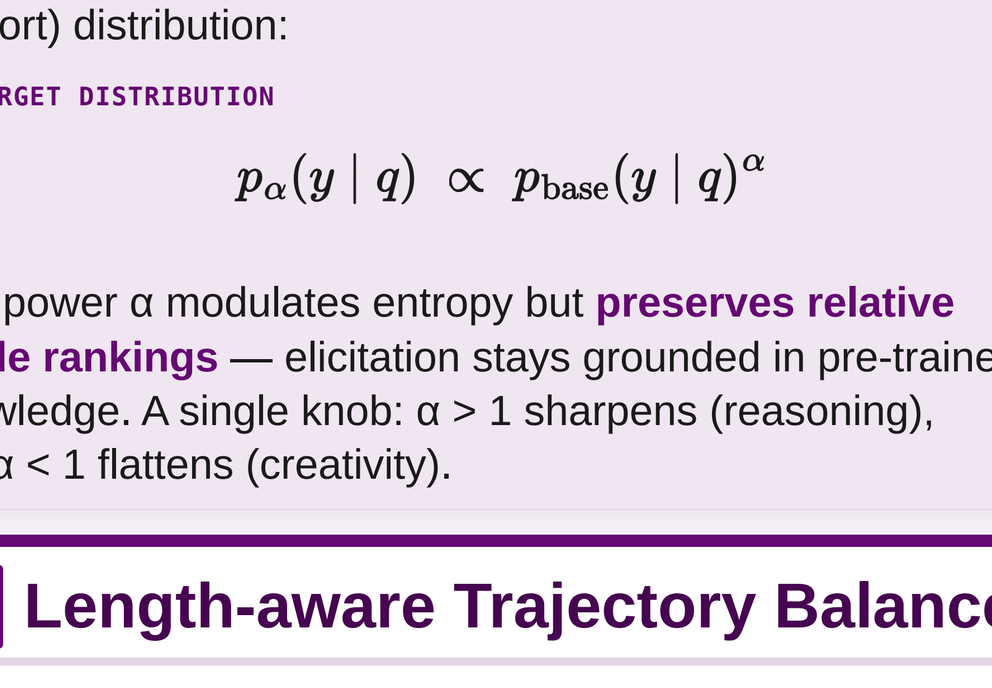
The power α modulates entropy but **preserves relative mode rankings** — elicitation stays grounded in pre-trained knowledge. A single knob: α > 1 sharpens (reasoning), 0 < α < 1 flattens (creativity).

02 Length-aware Trajectory Balance

GFlowNet gives an amortized sampler for the unnormalized target, but naive matching is dominated by length — collapse or repetition. The remedy: normalize the energy by $|y|$, a **length-normalized surface**:

THE OBJECTIVE

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



Training matches the α-power target on a length-normalized surface (clipped importance ratio + format penalty; no reward design). **Inference** stays standard decoding.

03 Sharpen (α > 1): reasoning

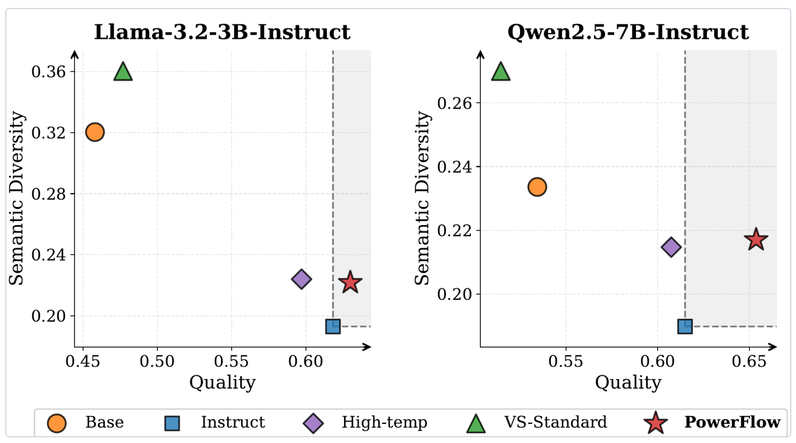
avg@16 over six benchmarks · Qwen2.5-Math-7B · α=4 · RLIF from official checkpoints.

Method	AIME25 ↑	AMC23 ↑	Avg ↑
EMPO	12.30	60.20	40.88
TTRL (best RLIF)	11.90	58.80	41.18
PowerFlow	**14.40**	**63.40**	**42.17**
GRPO (supervised)	12.90	63.40	42.38

Tops all unsupervised RLIF and **matches supervised GRPO** — beats it on 3 of 4 models, comparable on the 7B. It also keeps the most diverse solution strategies (AIME24 / 25): **4.05** > GRPO 3.93 > EMPO 3.80.

04 Flatten (0 < α < 1): creativity

Aligned instruct-tuned models · α=0.5 · quality vs. semantic diversity.



PowerFlow (**stars**) uniquely raises quality **while** lifting diversity (2 of 4 models shown).

05 Takeaway

- **One knob, two natures** — a single α reshapes the model's own distribution.
- **Principled + length-aware** — match the α-power target; no naive-matching length collapse.
- **Competitive, unsupervised** — rivals GRPO on reasoning; restores creativity.

Elicit, don't teach:
reshape what the model already knows.

PowerFlow · ICML 2026 · Corresponding author: Longbo Huang ✉ longbohuang@tsinghua.edu.cn
R. Chen · Y. Chen · Z. Li · L. Huang · IIIS, Tsinghua University · Code: github.com/Chenruishuo/PowerFlow