ICML International Conference On Machine Learning | 2026 POSTER

Unsupervised Capability Elicitation without Labels or External Rewards

GFLOWNET · ESCORT DISTRIBUTION MATCHING

清华大学 交叉信息研究院 Institute for Interdisciplinary Information Sciences, Tsinghua University

Code

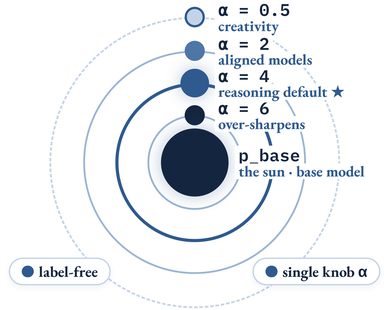
PRINCIPLED DISTRIBUTION MATCHING

PowerFlow

Unlocking the **dual nature** of LLMs — one control knob **α** sharpens the base distribution for reasoning or flattens it for creativity, with *no labels and no external rewards*.

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

Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University, Beijing, China · ✉longbohuang@tsinghua.edu.cn



Every orbit is an **α-power escort** of the base model, `pα(y|q) ∝ p_base(y|q)^α` — entropy is retuned while mode ranking is preserved.

`α > 1` → SHARPEN · `reasoning`
Tops every unsupervised baseline on all 4 models; **beats supervised GRPO on 3 of 4** (comparable on the 4th). Best unsupervised 7B avg **`42.17`**; solution diversity **`4.05`** > GRPO **`3.93`**.

`α < 1` → FLATTEN · `creativity`
The **only evaluated** method to raise quality *and* diversity together on aligned models — a Pareto-dominant shift over the Instruct baseline across four instruct models.

1 The Problem

Unsupervised RLIF elicits latent ability with **no labels and no external rewards** — but most methods hand-craft an intrinsic reward (self-certainty, entropy, majority voting). Lacking a principled optimization target, they are prone to the biases of their own reward, and under over-optimization drift into pathologies:

- Length **collapse** or repetitive **explosion**
- Overconfidence & majority-vote reward hacking
- **Mode collapse** — generative diversity is lost

Q. Can one principled objective elicit both reasoning and creativity — the LLM's dual nature — with no supervision?

2 Escort Target

PowerFlow reframes fine-tuning as **distribution matching**: pull the policy toward the base model's **α**-power (escort) distribution.

A-POWER ESCORT TARGET

$$p_\alpha(y \mid q) = \frac{p_{\text{base}}(y \mid q)^\alpha}{Z(q,\alpha)}$$

The escort retunes entropy while **preserving the base model's mode ranking and structure**. One knob sets the direction: **`α > 1`** sharpens toward latent reasoning paths; **`α < 1`** flattens the base distribution to recover the creativity typically suppressed during instruction tuning.

Self-derived from **`p_base`**, matching stays grounded in pre-trained knowledge; PowerFlow **surfaces latent capability**, not new skills.

3 Amortized Matching

The target's partition function is intractable, so PowerFlow amortizes it with a **GFlowNet**. The Trajectory-Balance objective is a variational surrogate for the reverse KL, training a policy that samples proportional to the unnormalized density.

Reasoning and creativity are then elicited under **standard single-pass decoding** — no inference-time MCMC, no verifier. PowerFlow amortizes the matching into training, so elicitation costs nothing extra at test time.

4 Length-Aware TB

Autoregressive log-probs scale with length, so naive matching chases length, not semantics. LA-TB projects the objective onto a **length-normalized energy surface**:

LENGTH-AWARE OBJECTIVE

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

Two guarantees: LA-TB is the **I-projection** (minimum-distortion length correction), and its global KL distortion is **second-order** in the length multiplier λ_q — so the induced warp stays small and vanishes as base lengths approach the calibrated budget.

~91% of α-power rankings preserved in practice (IR≈0.09).

The payoff is **monotonic, non-degenerate training**: trajectory-level matching collapses length and token-level matching decays via repetition, while PowerFlow holds both length and accuracy. A light format penalty (**ψ**, e.g. −0.5) and clip-higher importance sampling keep off-policy updates stable throughout training.

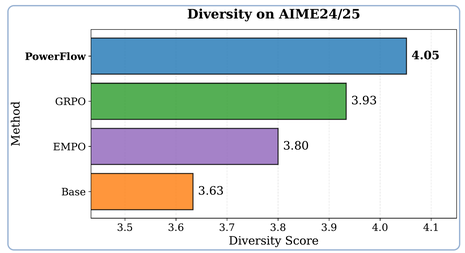
5 Reasoning

Average avg@16 over six math and science reasoning benchmarks (MATH500, Olympiad, AIME24/25, AMC23, GPQA).

Model	Base	Best RLIF	PowerFlow	GRPO supervised
Qwen2.5-1.5B	5.88	18.95 Intuitor	**19.85**	18.13
Qwen2.5-Math-1.5B	20.87	32.45 EMPO	**34.30**	32.75
Llama-3.2-3B-Inst	17.12	22.37 Intuitor	**22.88**	22.33
Qwen2.5-Math-7B	24.95	41.18 TTRL	**42.17**	**42.38**

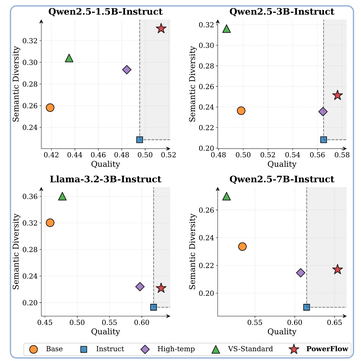
Tops **every unsupervised baseline on all four models**; beats GRPO on three, comparable on the 7B (**`42.17`** vs **`42.38`**).

6 Diversity



Sharpening need not collapse variety. On AIME24/25 (16 samples per problem, LLM-judged), PowerFlow keeps the widest strategy spread — **4.05** vs GRPO 3.93 and EMPO 3.80 — by rescaling the whole density instead of collapsing onto a single dominant reasoning mode.

7 Creativity



Quality vs. semantic diversity on aligned models, flattened with **`α = 0.5`**. PowerFlow (★) is the **only** method that moves up-and-right of the Instruct baseline — higher quality *and* more diversity — a Pareto-dominant shift across four instruct models.

8 Why It Works

Component ablation, Qwen2.5-Math-1.5B (avg@16).

Configuration	Avg	SEM
Dist. matching only (TB-traj)	≈ 0	—
+ format penalty only	14.30	0.27
+ length correction only	33.48	0.36
PowerFlow (full)	**34.30**	**0.30**

The **length-aware correction** carries the gain; without it, matching collapses toward 0. Format alone (7B) lands **5.95 below** PowerFlow.