CASE FILE — RLIF · α-POWER ICML 2026

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

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

Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University, Beijing, China

Correspondence: longbohuang@tsinghua.edu.cn

THE QUESTION ON THE BOARD

Does the base model already contain the reasoning we run RL to teach it?

An investigation into PowerFlow — principled distribution matching, no external supervision.

EXHIBIT A

The method on file

THE α-POWER (ESCORT) DISTRIBUTION

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

α > 1 · SHARPEN concentrates mass on latent reasoning paths.

α < 1 · FLATTEN releases suppressed creativity.

One knob: **α re-weights** the base model's own token probabilities, preserving relative rankings and mode structure (a monotonic reshaping), so it stays grounded in pre-trained knowledge — not heuristic rewards.

DEFAULTS α=4 base · α=2 instruct · α=0.5 creative

EXHIBIT B

No outside informants

PowerFlow matches p_α with a **GFlowNet** as an amortized sampler of the target, trained with a **length-aware Trajectory-Balance (LA-TB)** objective.

The catch. Log-probability falls roughly linearly with length $|y|$ — naive matching collapses it: trivially short when sharpening, repetitive long when flattening.

THE FIX — A LENGTH-AWARE ENERGY

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

Reparameterizing the partition function this way optimizes on a length-normalized surface — **~91%** of α-power rankings survive the correction (inversion rate ≈ 0.09).

≈ 0	33.48	34.30
avg when the length fix is dropped	recovered by the fix alone	full PowerFlow (Qwen2.5-Math-1.5B)

EXHIBIT C

The headline evidence

avg@16 over 6 benchmarks — MATH500 · OlympiadBench · AIME24 · AIME25 · AMC23 · GPQA-diamond

Model	Best RLIF baseline (unsup.)	PowerFlow (ours, α > 1)	GRPO (supervised)
Qwen2.5-1.5B	Intuitor 18.95	**19.85**	18.13
Qwen2.5-Math-1.5B	EMPO 32.45	**34.30**	32.75
Llama-3.2-3B-Instruct	Intuitor 22.37	**22.88**	22.33
Qwen2.5-Math-7B	TTRL 41.18	42.17	**42.38**

PowerFlow tops **every** unsupervised RLIF baseline on all 4 models. It beats **supervised** GRPO on 3 of 4 (all > 1σ) and is comparable on Qwen2.5-Math-7B (42.17 vs 42.38) — with **no labels and no verifier**.

CROSS-EXAMINATION

Alternative explanations, ruled out

1 "Just lower the temperature?" RULED OUT

Temperature scaling *averages* future likelihoods; the power distribution p^α *sharpens* high-likelihood paths — a different target. The Low-temp baseline reaches only **35.93** vs PowerFlow **42.17** (Qwen2.5-Math-7B).

2 "Just format compliance?" RULED OUT

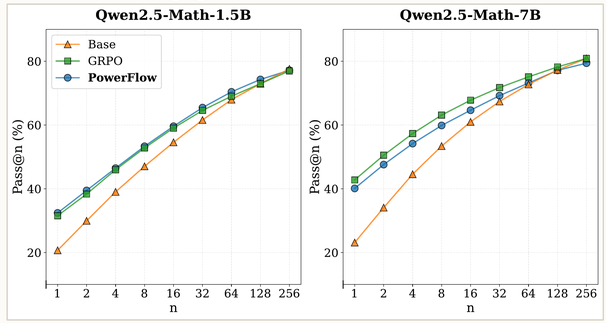
A format-only variant (α = 1) reaches **36.22** — **5.95** points below PowerFlow (42.17) on the same 7B suite, well past 1σ.

3 "Injecting new knowledge?" RULED OUT

Pass@256 shows the Base model catching up (Exhibit D): the capability was latent — PowerFlow raises **elicitation efficiency**.

EXHIBIT D

The smoking gun

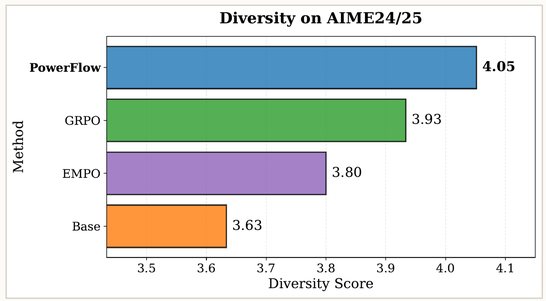


Pass@n on OlympiadBench. As n grows, the **Base** model catches up (by pass@256) — the reasoning was **already in the model**. PowerFlow surfaces it more efficiently; it doesn't add new knowledge.

EXHIBIT E

Strategies intact

Sharpening for accuracy does not cost variety: PowerFlow keeps **superior solution-strategy diversity**, unlike RLIF methods that collapse toward one mode.

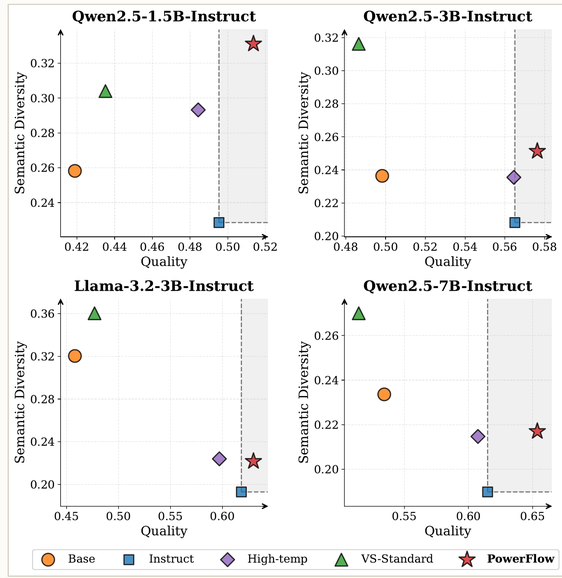


Strategy diversity, AIME24/25 (LLM-judge: DeepSeek-V3.2, 16 samples per problem): PowerFlow 4.05 > GRPO 3.93 > EMPO 3.80 > Base 3.63.

EXHIBIT F

The other face

Flip the knob to **α = 0.5**: PowerFlow flattens the distribution, undoing the over-sharpening ("typicality bias") that alignment training bakes into the model.



Quality vs. semantic diversity, 4 instruct models. PowerFlow (★) lands in the upper-right Pareto-improvement region — more diverse **and** higher quality than the Instruct baseline (■).

Across the 4 models it is the **only** method to improve **both** semantic diversity **and** quality over the Instruct baseline at once. High-temp buys diversity by losing quality; VS-Standard gives up quality at model scales ≤ 7B.

One knob, dual nature: α > 1 surfaces latent reasoning; α = 0.5 releases suppressed creativity.

THE VERDICT

The reasoning is already inside the base model. Re-weight its own probabilities toward $p_\alpha \propto p_{\text{base}}^{\alpha}$ — no labels, no verifier — and it surfaces; push α below 1 and creativity does too.

42.17	3 / 4	α>1 · α<1
best unsupervised avg@16, Qwen2.5-Math-7B (TTRL 41.18 · GRPO 42.38)	tested models beat supervised GRPO — with zero labels	sharpen · flatten — one knob, dual nature

Case file →
github.com/Chenruishuo/PowerFlow