Beyond Outcome: Trajectory-Driven Prompt Optimization via Multi-Dimensional Rewards

Zhixiang Liang[1,2,3*] Yixiang Huang[1*] Chenjing Cai[1†] Hao Wu[1] Peihang Liu[1] Qi Liu[1,4] Qiqi Zhong[1,5] Jiuchong Gao[1†] Jinghua Hao[1] Renqing He[1]

[1] Meituan [2] School of Artificial Intelligence, University of Chinese Academy of Sciences [4] Peking University [5] UIBE

[3] Beijing Key Lab of Molecular Imaging Technology Research and Translation, CAS Key Lab of Molecular Imaging, CASIA

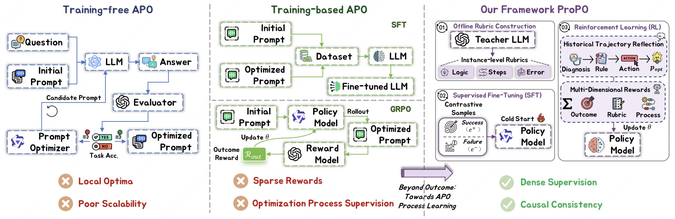
NEURAL INFORMATION PROCESSING SYSTEMS NeurIPS 2026

Research Motivation 01

- Most APO rewards the final answer.

- The optimization process stays unsupervised.

Training-free: local optima | Outcome-only RL: sparse rewards

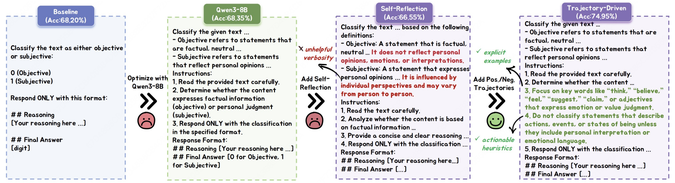


ProPO learns how to derive better prompts.

Key Evidence 02

Trajectory-grounded reflection

Execution examples expose errors and guide revisions.



Self-reflection: 66.55% → Trajectory reflection: 74.95%

Precise rubric supervision

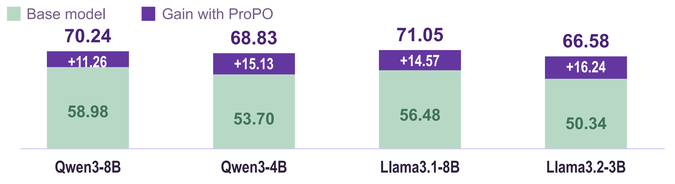
Instance-level criteria provide denser guidance.

Rubric strategy	Avg. (%)	Gain (pp)
Without rubric	66.54	-
Manually written	65.73	-0.81
Dataset-level	66.61	+0.07
Instance-level (ProPO)	**70.24**	**+3.70**

+3.70 pp average gain, including +8.89 pp on AIME.

Gains across policy backbones

Works across model families and parameter scales.



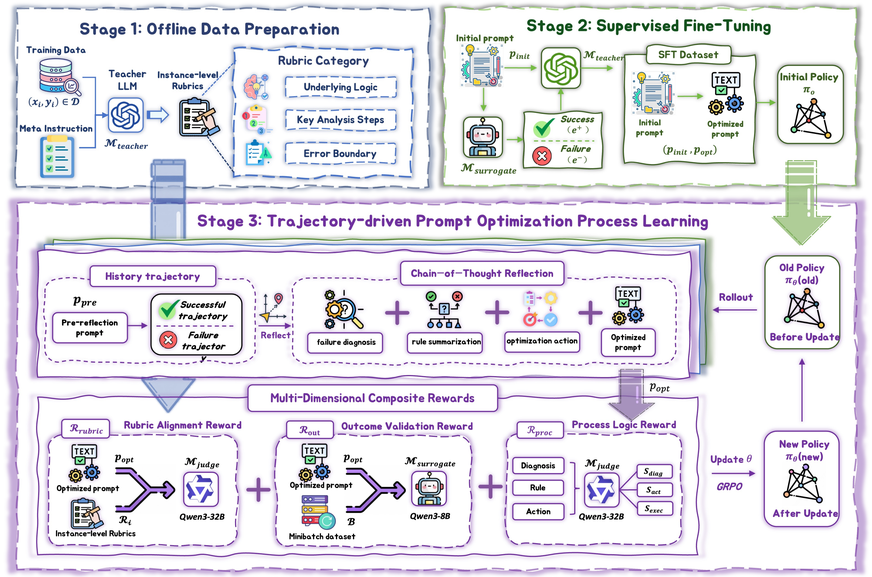
Average gains: +11.26 to +16.24 pp.

ProPO Method 03

Trajectory-driven process learning

Successful and failed executions ground structured reflection.

Offline preparation → SFT cold start → GRPO process learning



Policy / surrogate: Qwen3-8B Reward judge: Qwen3-32B Teacher: GPT-4.1

Multi-dimensional rewards

Three complementary rewards supervise each trajectory.

Outcome
Does the prompt work?
Validation-minibatch accuracy

Adaptive rubric
Does it honor the task?
Instance-level criteria scored by an LLM judge

Process logic
Is reasoning coherent?
Diagnosis, action and execution consistency

$$R_{total} = R_{out} + R_{rubric} + R_{proc}$$

Dense semantic feedback and explicit process supervision.

Causal consistency checks

All three binary checks must pass for the process reward.

Correct diagnosis → Action fixes the error → Prompt executes the action

Process reward = 1 for a consistent chain; otherwise 0.

Experimental Results 04

Main performance

Best average accuracy across six benchmarks.

70.24% average accuracy | **+6.47 pp** vs. outcome GRPO

Mean ± std over seeds 1, 42, 2026. Higher is better.

Method	IFBench	AIME	LBench	SST-5	Subj	TREC	Avg.
Training-free APO							
TextGrad	36.73±0.28	18.89±1.57	59.53±0.62	54.34±0.28	68.70±0.04	85.07±0.93	53.88
SPO	25.06±3.54	55.55±3.14	57.63±0.05	54.74±0.48	72.93±0.20	72.60±0.82	56.42
Trace	26.08±1.25	57.78±1.57	57.13±0.17	52.46±0.26	73.85±0.40	76.60±0.43	57.32
GEPA	36.39±1.69	57.78±3.14	55.73±2.25	53.51±0.44	80.18±0.80	86.00±0.33	61.60
Training-based APO / baselines							
Zero-shot	33.55±0.97	57.78±1.57	54.67±0.26	54.31±0.51	68.17±0.08	85.40±0.33	58.98
SFT	37.18±0.43	58.89±1.57	58.13±0.69	54.13±0.80	70.37±0.15	88.40±0.43	61.18
GRPO (outcome)	38.08±0.50	57.78±5.67	58.70±1.72	55.93±0.32	82.42±0.54	89.73±0.09	63.77
StablePrompt	27.66±0.58	58.89±1.57	62.10±1.56	54.88±0.72	77.63±0.41	87.87±0.25	61.51
PromptR1	38.66±0.32	63.30±4.77	58.27±0.61	55.76±0.07	79.90±0.11	77.40±0.85	62.22
ProPO	**40.59±0.43**	**71.11±1.57**	**63.93±0.78**	**58.71±0.46**	**93.87±0.52**	**93.20±0.29**	**70.24**

+6.47 pp over outcome-only GRPO on average.

LBench = LiveBench-Math. ProPO row highlighted.

Ablation studies

Complementary rewards and consistent process checks.

Reward components

Accuracy (%); drops vs. full.

Setting	IFB	AIME	LBench	SST-5	Subj	TREC	Avg.	Drop (pp)
Outcome only	38.08	57.78	58.70	55.93	82.42	89.73	63.77	-6.47
w/o outcome	37.53	57.78	57.80	55.17	87.27	91.27	64.47	-5.77
w/o rubric	38.20	62.22	58.20	57.51	91.20	91.93	66.54	-3.70
w/o process	39.68	62.22	61.13	57.70	91.30	92.07	67.35	-2.89
ProPO (full)	**40.59**	**71.11**	**63.93**	**58.71**	**93.87**	**93.20**	**70.24**	**-**

Internal process consistency checks

Setting	IFB	AIME	LBench	SST-5	Subj	TREC	Avg.	Drop (pp)
w/o execution	38.88	71.11	64.20	58.26	94.08	93.00	69.92	-0.32
w/o action	38.88	68.89	62.07	58.04	93.08	92.47	68.91	-1.33
w/o diagnosis	37.99	65.56	61.80	57.92	91.88	92.00	67.86	-2.38
w/o process	39.68	62.22	61.13	57.70	91.30	92.07	67.35	-2.89
ProPO (full)	**40.59**	**71.11**	**63.93**	**58.71**	**93.87**	**93.20**	**70.24**	**-**

Largest single-check drop: -2.38 pp without diagnosis.

Accuracy and runtime

Total across six benchmarks. RL methods use 8×A100.

Method	Total runtime (hours, lower is better)	Avg. acc.
TextGrad	2.46 h	53.88%
SPO	3.07 h	56.42%
Trace	7.94 h	57.32%
GEPA	1.08 h	61.60%
StablePrompt	14 h	61.51%
PromptR1	29 h	62.22%
ProPO @20 steps	2.6 h	**66.58%**
ProPO @80 steps	12 h	**70.24%**

70.24% in 12 h; 58.6% less time and +8.02 pp vs. PromptR1.

* Equal contribution. † Corresponding authors. Contact: liangzhixiang2025@ia.ac.cn