DuPO: Enabling Reliable Self-Verification via Dual Preference Optimization

Shuaijie She[♠♡], Yu Bao[♡], Yu Lu[♡], Lu Xu[♡], Tao Li[♡], Wenhao Zhu[♡],
Jianbing Zhang[♠], Shujian Huang[♠†], Shanbo Cheng[♡†], Lu Lu[♡], Yuxuan Wang[♡]

[♠]Nanjing University [♡]ByteDance Seed

TL;DR *We introduce DuPO, a dual learning-based preference optimization framework that generates reliable self-supervised rewards via generalized task duality across diverse tasks including mathematical reasoning (+6.4 avg. points) and multilingual translation (+2.1 COMET over 756 directions).*

PAPER

Background & Motivation

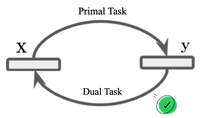
RLHF relies on inconsistent and costly human annotations | **RLVR is limited to verifiable tasks (e.g., math, code)**

⇧ Solve with self-verification ⇧

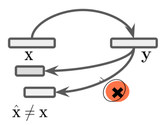
Dual learning proposes the dual task to provide reward

⇩ Face two challenges ⇩

Challenge I: Limited Duality in Non-Invertible Tasks | **Challenge II:** Bidirectional Competence Asymmetry

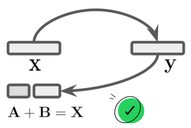


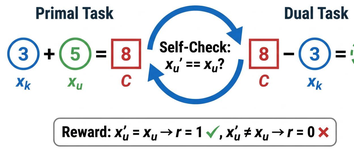
Traditional dual learning requires strict duality



E.g., 3 + 5 → 8 but 8 → 10-2 / 9-1…

E.g., Good at X→En but Weak at En→X



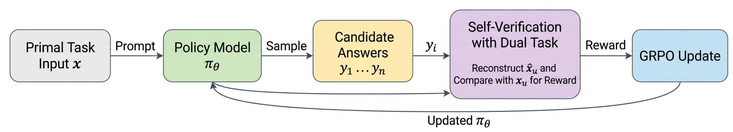


DuPO: Duality-based Preference Optimization

- **Input Decomposition:** split the input x into a known component x_k and an unknown component x_u.
- **Complementary Dual Task:** recovers x_u from the primal output y and x_k, serving as a self-supervised reward signal

$$\mathcal{T}_{cd} : (\mathbf{y}, \mathbf{x}_k) \mapsto \hat{\mathbf{x}}_u. \qquad \forall \mathbf{x} \in \mathcal{X},\ \mathbf{y} = \mathcal{T}_p(\mathbf{x}) :\ d(\mathbf{x}_u, \mathcal{T}_{cd}(\mathbf{y}, \mathbf{x}_k)) \le \epsilon,$$

- **Generalized Duality Reward:** $r(\mathbf{x}, \mathbf{y}) \propto \exp\left(-\lambda \cdot d\left(\mathbf{x}_u, \mathcal{T}_{cd}(\mathbf{y}, \mathbf{x}_k)\right)\right),$
- **Policy Optimization:** $\mathcal{J}(\theta) = \mathbb{E}_{\mathbf{y} \sim \pi_\theta(\mathbf{y}|\mathbf{x})}\left[r(\mathbf{x}, \mathbf{y})\right]$



How to Prepare Dual Task? Automatic Pipeline

Generation:
- Regex for variable extraction/masking

Unknown-component Selection:
- Answerability of the dual question
- Uniqueness of the correct completion

Select x_u that robustly satisfies the properties of duality without accessing annotations.

Examples in AIME24 and Flores

	Scenario 1: DuPO on Mathematical Reasoning
Primal Task	Let $\triangle ABC$ have circumcenter O and incenter I with $\overline{IA} \perp \overline{OI}$, circumradius **13**, and inradius **6**. Find $AB \cdot AC$. (Correct Answer: 468)
Dual Task #1	Let $\triangle ABC$ have circumcenter O and incenter I with $\overline{IA} \perp \overline{OI}$, circumradius V_{sk}, and inradius 6. Find $AB \cdot AC$. Check your work: If the solution for above question is $\boxed{answer}$, what must V_{sk} have been?
Dual Task #2	Let's examine: Let $\triangle ABC$ have circumcenter O and incenter I with $\overline{IA} \perp \overline{OI}$, circumradius 13, and inradius V_{rj}. Find $AB \cdot AC$. When the solution for above question is $\boxed{answer}$, what's the corresponding V_{rj}?
Candidates	**Answer: 468 Backward Accuracy: 69.1%** **Answer: 108 Backward Accuracy: 0%** **Answer: 312 Backward Accuracy: 0%**
	Scenario 2: DuPO on Machine Translation (MT)
Primal Task	Translate to Chinese: As knowledge of Greek declined, the West found itself cut off from its Greek philosophical and scientific roots.
Reference	随着希腊知识的衰落，西方脱离了其希腊哲学和科学根源。
Primal MT #1	随着希腊语知识的衰落，西方发现自己与希腊的哲学和科学根源失去了联系。**(BLEU: 45.85)**
Dual MT #1	As knowledge of Greek declined, the West found itself cut off from its philosophical and scientific roots in Greece.**(BLEU: 82.07)**
Primal MT #2	随着对希腊语的了解逐渐消失，西方发现自己与希腊哲学和科学根源隔绝开来。**(BLEU: 28.65)**
Dual MT #2	As understanding of the Greek language gradually fades, the West finds itself cut off from the roots of Greek philosophy and science.**(BLEU: 16.11)**

Improve LLM Performance across Diverse Tasks without Labels (Math Reasoning and Multilingual Translation)

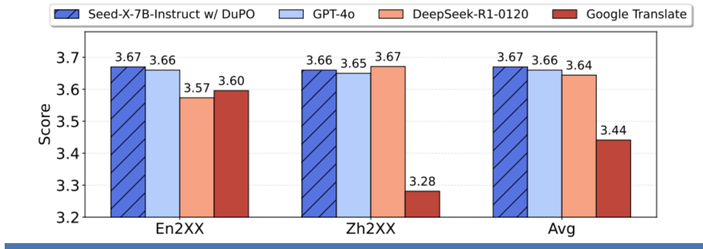
DuPO consistently boosts math reasoning across model scales and benchmarks, yielding an average gain of 6.4 points, outperforming commercial LLMs including Sonnet4-Thinking and DeepSeek-R1-0120.

Model	AMC23	AIME24	AIME25	HMMT	Avg.
DeepSeek-R1-0120	97.7	79.8	70.0	44.2	72.9
Claude-Sonnet4-Thinking	97.5	82.5	70.0	48.3	74.6
Doubao-1.5-Thinking	99.4	86.3	73.3	57.7	79.2
Doubao-1.6-Thinking	98.8	88.4	83.4	60.1	82.7
DeepSeek-R1-0528	99.4	91.4	87.5	71.4	87.4
DeepSeek-R1-Distill-Qwen-1.5B	67.5	20.0	20.0	13.3	30.2
w/ DuPO (ours)	72.5	30.0	26.7	16.7	36.5 (+6.3)
DeepSeek-R1-Distill-Qwen-7B	85.0	56.7	36.7	20.0	49.6
w/ DuPO (ours)	90.0	63.3	40.0	26.7	55.0 (+5.4)
Qwen3-4B	95.0	70.0	66.7	40.0	67.9
w/ DuPO (ours)	97.5	83.3	70.0	46.7	74.4 (+6.5)
OpenReasoning-Nemotron-7B	95.0	83.3	73.3	56.7	77.1
w/ DuPO (ours)	97.5	83.3	90.0	66.7	84.4 (+7.3)

Model	BLEU	COMET	BLEURT	Avg.
Qwen3-8B	21.7	84.8	65.8	57.4
Doubao-1.5-Thinking	26.2	87.9	71.7	61.9
Qwen3-235B-22B	28.4	88.8	73.9	63.7
DeepSeek-R1-0528	30.2	89.2	75.0	64.8
Seed-X-7B-Instruct	28.8	87.0	72.6	62.8
w/ DuPO (ours)	30.3	89.1	74.6	64.7

DuPO improves Seed-X-7B by +2.1 COMET across 756 translation directions, on par with DeepSeek-R1.

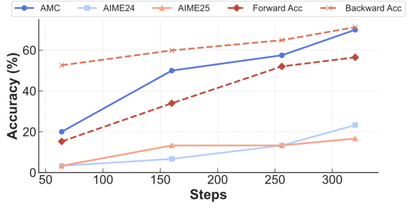
Human evaluation confirms our performance matches GPT-4o and surpasses Google Translate.



Scale to Various Backbones Effectively

DuPO improves LLaMA-3.1, OctoThinker, and even Qwen3-4B-Base, demonstrating scalability and robustness to varying initial capabilities (also evidenced by co-evolving performance on both primal and dual tasks).

Model	AMC23	MATH500	Avg.
LlaMA-3.1-8B	2.5	13.6	8.1
w/ SimpleRL-Zoo	15.0	23.0	19.0
w/ DuPO (ours)	20.0	44.2	32.1
OctoThinker-8B-Hybrid-Base	5.0	42.6	23.8
w/ DuPO (ours)	55.0	70.0	62.5



Scale Reasoning during Inference without Training

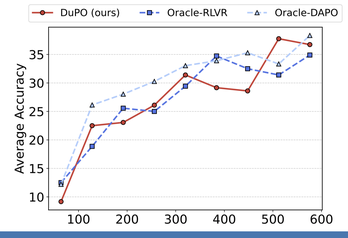
Model	AIME24	AIME25	Avg.
DeepSeek-R1-0120	79.8	70.0	74.9
Claude-Sonnet4-Thinking	82.5	70.0	76.3
DeepSeek-R1-Distill-1.5B	20.0	20.0	20.0
w/ DuPO rewarding	53.3	24.1	38.7 (+18.7)
Qwen3-4B	70.0	66.7	68.4
w/ DuPO rewarding	86.6	68.9	77.7 (+9.3)

DuPO acts as a training-free reranker

+18.7 on Distill-1.5B and +9.3 on Qwen3-4B

surpassing DeepSeek-R1 and Sonnet4-Thinking.

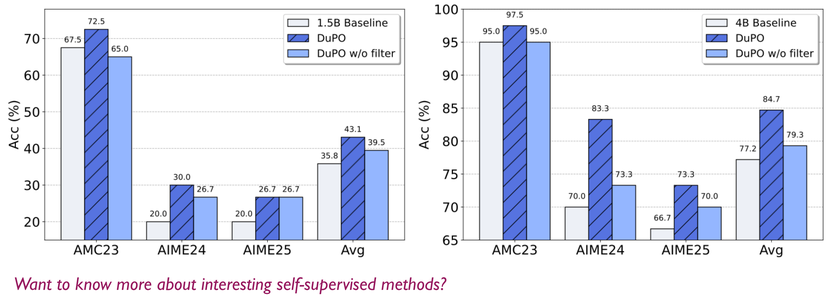
Achieve Similar Performance with Oracle



Without using any ground truth labels, DuPO closely tracks Oracle Baselines' accuracy curve, achieving comparable performance using self-supervised rewards.

Better Task Duality Leads to Improved Performance

Removing the unknown-component selection strategy degrades avg. accuracy by 3.6–5.4 pts, validating its role in providing cleaner reward signals.



Want to know more about interesting self-supervised methods?

[Factuality] CoP: Factual Inconsistency Detection by Controlling the Preference. In AAAI 2023.
[Reasoning] Advancing Multilingual Reasoning through Multilingual Alignment-as-Preference Optimization. In ACL 2024.
[LongContext] Improving Long-Context Translation via Self-Supervised Dual Learning. In ACL 2026.