# Turning Structure into Reward: Toward Reliable Self-Supervised Reinforcement Learning

Shuaijie She
shesj@smail.nju.edu.cn

**TL;DR** Our research aims to develop *general-purpose self-supervised RL* methods that scale across diverse tasks while *reducing reliance on annotations and noisy synthetic data.*

Our central idea is to turn structures already present in tasks and models—including task duality and uneven model capabilities—into reward and preference signals. Through DuPO and MAPO, we investigate how such structural feedback can support reliable self-improvement across reasoning and generation.

PROFILE

## 1. Background: Why Self-Supervision?

**Human Feedback**
High annotation costs
Potential annotator bias

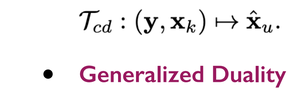
**LLM-as-Judge**
Rely on stronger teacher
Prone to errors and bias

**Verifiable Rewards**
Requires task-specific verifiers
Limited task coverage

Teaching models every task by hand does not scale.
We need scalable and reliable rewards

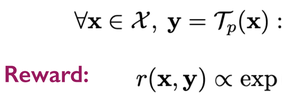
Where can self-supervision come from?

**Task duality structure**



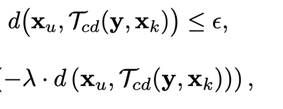

**Reconstruct to verify**
DuPO · More general tasks

**Capability structure**




**Align to transfer**
MAPO · Reasoning ability transfer

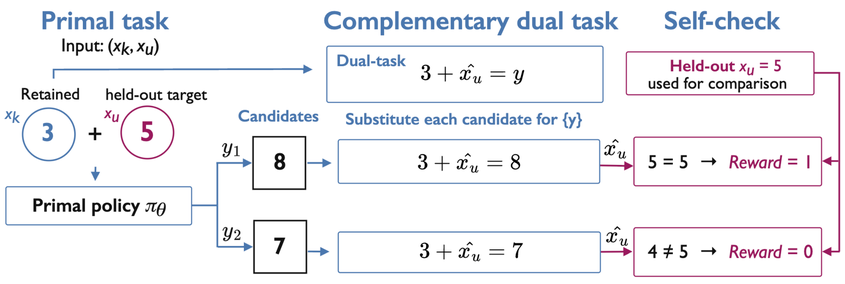
**Input–output structure**




**Prob Diff Probes factuality**
CoP · Factuality evaluation

## 2. DuPO: Generalized Task Duality

- **Two Roles:** Primal task produces $y$ conditioned on input $x$; Dual task uses $y$ to reconstruct the input $x$; and assess its quality. (e.g., Zh → En → Zh)
- **Feedback Mechanism:** Incorrect or information-losing $y$ → larger mismatch with the original input $x$ → lower reward.
- **Scaling Challenge:** Most tasks lack primal–dual symmetry. (e.g. 3 + 5 → 8 but 8 → 10-2 / 9-1… ; Good at Bn→En but Weak at En→Bn)




- **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

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

- **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*

**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.

***Extends to more general tasks without annotations, with automatic, reliable dual-task generation and verification.***

**INSTANTIATIONS / What is reconstructed?**

**Long-context Input/Output** — LongDu

Target: the full source document. Retained source content: none ($x_k = \varnothing$).

**Long Context Input** → **Long Output** → **Reconstruction of Input**

Score reconstruction, with target-language and fluency checks to discourage shortcuts.

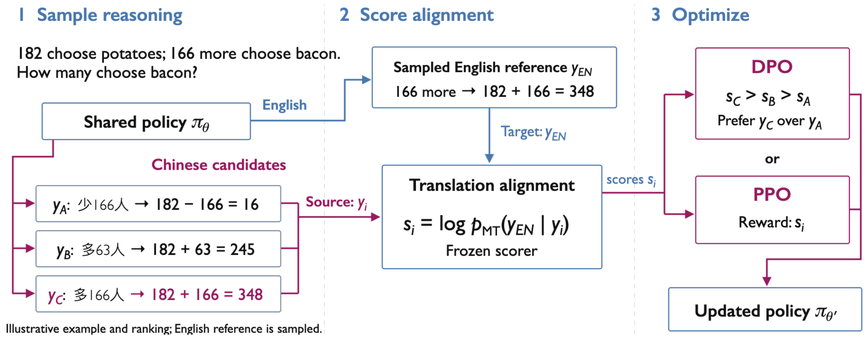
**Possible extension: Multimodality, Agentic RL, Trading System?**

Reconstruction mismatches can reveal reasoning or execution errors, providing *self-supervised feedback that reduces reliance on annotations and noisy synthetic data.*

**Report / Factor Specifications** → **Summary / Decision** → **Code / Action** → **Recover Key Facts** → **Recover Constraints**

## 3. MAPO: Alignment for Strong-to-Weak Transfer

- **Uneven Capabilities:** LLMs excel at some tasks but struggle with others.
- **Existing Methods:** costly annotations, noisy synthetic data, and poor generalization.




- **Choose and sample:** Identify a stronger task and sample the same model's outputs on both the strong and weak tasks.
- **Align and train:** Use a cross-task aligner to measure output alignment, then train the policy model to favor weak-task responses that align with strong-task outputs.
- ***No need for reasoning annotations; Better generalization.***

**INSTANTIATIONS / What is aligned?**

**Possible extension: Code/Math** — Possible extension

A bridge model (e.g., an LLM) scores alignment for cross-task capability transfer between experts

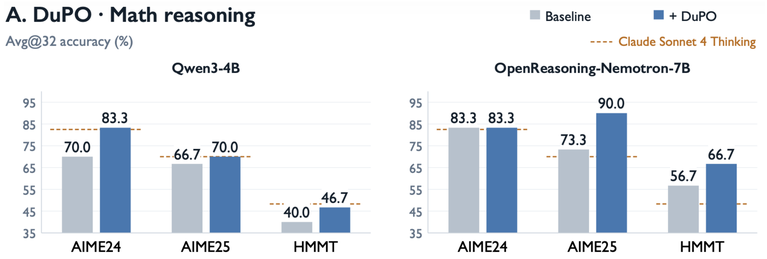
**Possible extension on Trading System?** Cross-Asset Reasoning Transfer

Transfer reasoning across related assets through shared factors and constraints.

**Reference Reasoning (Asset A)** → **Alignment Score** ← **Candidate Reasoning (Asset B)**

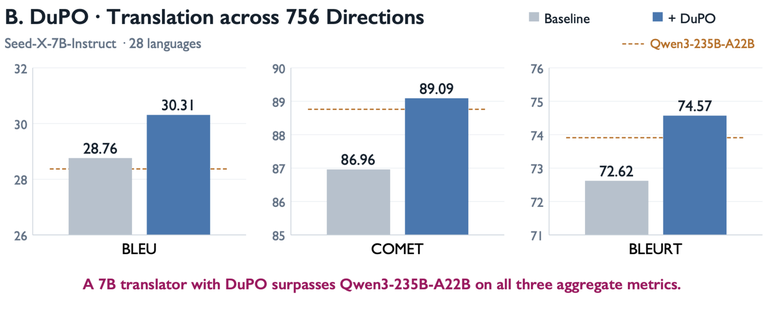
## 4. Evaluation: Performance and Generalization

**A. DuPO · Math reasoning**



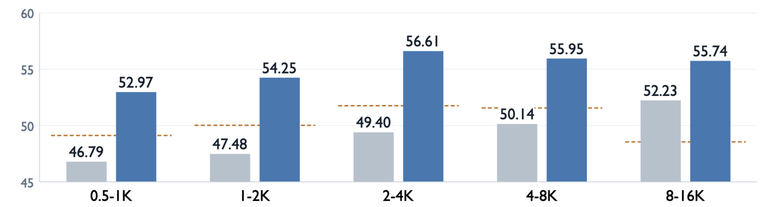

DuPO improves both backbones: mean gains of +7.8 pp on Qwen3-4B and +8.9 pp on Nemotron-7B.

**B. DuPO · Translation across 756 Directions**



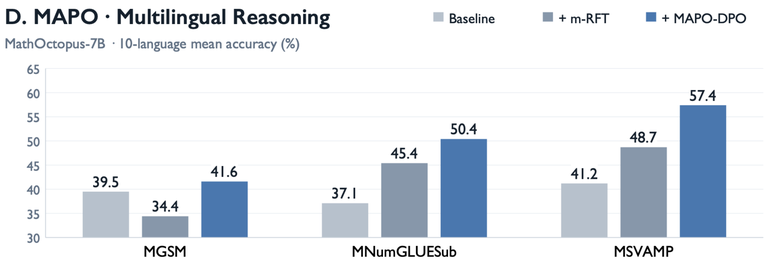

A 7B translator with DuPO surpasses Qwen3-235B-A22B on all three aggregate metrics.

**C. LongDu · Long document translation**




Post-trained on 0.5-2K-token documents, LongDu beats DeepSeek-V3.1 in BLEU through 16K.

**D. MAPO · Multilingual Reasoning**




MAPO avoids MGSM forgetting (m-RFT: −5.1 pp) and generalizes to unseen MSVAMP (+16.2 vs +7.5 pp).

*Want to know more about interesting self-supervised method?*

[General] DuPO: Enabling Reliable LLM Self-Verification via Dual Preference Optimization. In ICLR
[LongContext] Improving Long-Context Translation via Self-Supervised Dual Learning. In ACL.
[Reasoning] Advancing Multilingual Reasoning through Multilingual Alignment-as-Preference Optimization. In ACL
[Factuality] CoP: Factual Inconsistency Detection by Controlling the Preference. In AAAI.