
000 001 002 003 004 005 006 007 008 009 010 DUPO: ENABLING RELIABLE SELF-VERIFICATION VIA DUAL PREFERENCE OPTIMIZATION

005
006
007
008
009
010
Anonymous authors
006
007
008
009
010
Paper under double-blind review

ABSTRACT

011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
We present DuPO, a dual learning-based preference optimization framework that
generates annotation-free feedback via the generalized duality. DuPO addresses
two key limitations: Reinforcement Learning with Verifiable Rewards (RLVR)'s
reliance on costly labels and applicability restricted to verifiable tasks, and tra-
ditional dual learning's restriction to strictly dual task pairs (e.g., translation
and back-translation). Specifically, DuPO decomposes a primal task's input into
known and unknown components, then constructs its dual task to reconstruct the
unknown part using the primal output and known information (e.g., reversing math
solutions to recover hidden variables), broadening applicability to non-invertible
tasks. The quality of this reconstruction serves as a self-supervised reward to opti-
mize the primal task, synergizing with LLMs' ability to instantiate both tasks via a
single model. Empirically, DuPO achieves substantial gains across diverse tasks:
it enhances the average translation quality by 2.1 COMET over 756 directions,
boosts the mathematical reasoning accuracy by an average of 6.4 points on three
challenge benchmarks, and enhances performance by 9.3 points as an inference-
time reranker (trading computation for accuracy). These results position DuPO as
a scalable, general, and annotation-free paradigm for LLM optimization.

1 INTRODUCTION

027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
Large Language Models (LLMs) (Yang et al., 2025; Grattafiori et al., 2024; Liu et al., 2024; Ope-
nAI, 2023; AI, 2024; DeepMind, 2025) have shown remarkable progress in tasks like mathemat-
ical reasoning (Yang et al., 2024b; Shao et al., 2024; Chen et al., 2025) and multilingual transla-
tion (Cheng et al., 2025; Zhu et al., 2024; Li et al., 2024b;c). To further enhance these capabilities,
researchers have increasingly adopted reinforcement learning (RL) paradigms like Reinforcement
Learning from Human Feedback (RLHF) (Grattafiori et al., 2024; Liu et al., 2024; Yang et al., 2025)
and Reinforcement Learning with Verifiable Rewards (RLVR) (DeepSeek-AI, 2025; Team, 2025; Yu
et al., 2025; He et al., 2025; Hu et al., 2025b) have gained traction. Specifically, RLHF aligns mod-
els with human preferences but relies on costly, inconsistent human annotations (Lee et al., 2023;
Zhang et al., 2024). RLVR addresses this for objective tasks (e.g., math, code) via binary rewards
from verifiable answers, reducing annotation burdens. However, RLVR still depends on external
supervision: acquiring verifiable answers remains a bottleneck, limiting scalability. Moreover, it
struggles with generative tasks (e.g., multilingual translation), where single references cannot cap-
ture diverse high-quality outputs (Jia et al., 2025; Callison-Burch et al., 2006). Recent attempts (e.g.,
AI-Feedback/RLAIF (Lee et al., 2023), Constitutional AI (Bai et al., 2022)) merely swap dependen-
cies (human labels → teacher models/rules), failing to resolve the core bottleneck.

045
046
047
048
049
050
Dual learning (He et al., 2016) offers a self-supervised alternative by leveraging *task duality* to
generate intrinsic feedback: through paired “primal” and “dual” tasks (e.g., translation and back-
translation (Sennrich et al., 2015)), models validate outputs via cycle consistency, eliminating re-
liance on external labels. Given that LLMs possess diverse capabilities from extensive pretraining,
they could be trained across various tasks. However, applying this framework to LLMs is non-trivial,
which faces two critical challenges:

051
052
053
1. **Limited Duality in Mutually Non-Invertible Tasks:** Most real-world LLM tasks (e.g.,
creative writing, math reasoning) lack strict invertibility. LLM's output (e.g., a math solu-
tion) rarely contains enough information to reconstruct its input (e.g., the original problem),
breaking the duality cycle.

054 2. **Bidirectional Competence Asymmetry:** LLMs often exhibit uneven performance across
055 primal/dual tasks (e.g., strong at solving math problems but weak at generating problems
056 from solutions). Noisy self-signals from asymmetric tasks hinder optimization, reducing
057 the framework’s utility.

058 These mismatches render traditional dual learning ill-suited for general LLM optimization, leaving
059 it an open challenge.

061 In this paper, we propose **DuPO (Dual Learning-based Preference Optimization)**, a framework
062 that aligns LLM generalization with a (relaxed) duality applicable to general tasks. At its core lies
063 a *generalized duality framework* (§3.2) built on *complementary relationships*: it decomposes each
064 input x into disjoint known (x_k) and unknown (x_u) components, then designs the dual objective to
065 reconstruct only x_u from the primal output y and the known input x_k , rather than inverting the full
066 input. This framework resolves two asymmetries: it restores sufficient information flow between
067 the primal and dual tasks (task asymmetry) and reduces the complexity burden on the dual task side
068 (capability asymmetry). The formulation naturally synergizes with LLMs: their broad foundational
069 capabilities allow a single model to instantiate both primal and dual tasks without specific architec-
070 tures, while the dual task converts the model’s outputs into self-supervised reward signals, enabling
071 continual improvement without external annotations. This bidirectional benefit addresses a critical
072 challenge in LLM development: obtaining high-quality feedback for capability enhancement.

073 We empirically validate DuPO on two representative tasks: mathematical reasoning and multilingual
074 translation, demonstrating significant and consistent improvements. By applying DuPO to one of the
075 strongest translation LLM, Seed-X-7B-Instruct (Cheng et al., 2025), we demonstrate a significant
076 performance gain of 2.1 COMET points on the multilingual translation benchmark, bringing the
077 7B model to performance comparable to ultra-large SOTA systems. In mathematical reasoning, our
078 method yields robust gains across models of varying scales, from 1.5B to 7B parameters; notably,
079 DuPO improves the Qwen3-4B (Yang et al., 2025) model’s score on three challenging mathematical
080 benchmark by 6.4 percentage points. Our comprehensive ablation studies confirm that our design,
081 the generalized duality, is crucial for achieving these results. Beyond training, DuPO acts as a
082 *reranking mechanism* at inference, boosting performance by 9.3 points without finetuning—enabling
083 smaller models to outperform stronger ultra-large LLMs like DeepSeek-R1 (DeepSeek-AI, 2025)
084 even without training. In summary, DuPO reimagines task duality for non-invertible LLM tasks.
085 It eliminates external annotation reliance, scales across tasks/domains, and enhances both training
086 and inference—offering a scalable path to align LLMs with diverse goals using self-supervised
087 feedback, **marking a promising first step toward unlocking self-verification for broader, open-ended
088 domains.**

088 2 PRELIMINARIES

090 We can cast various tasks as conditional generation: with input space \mathcal{X} and output space \mathcal{Y} , the
091 model generates $y \in \mathcal{Y}$ via the LLM $\pi_\theta(y | x)$ given an input $x \in \mathcal{X}$.

092 2.1 PREFERENCE OPTIMIZATION

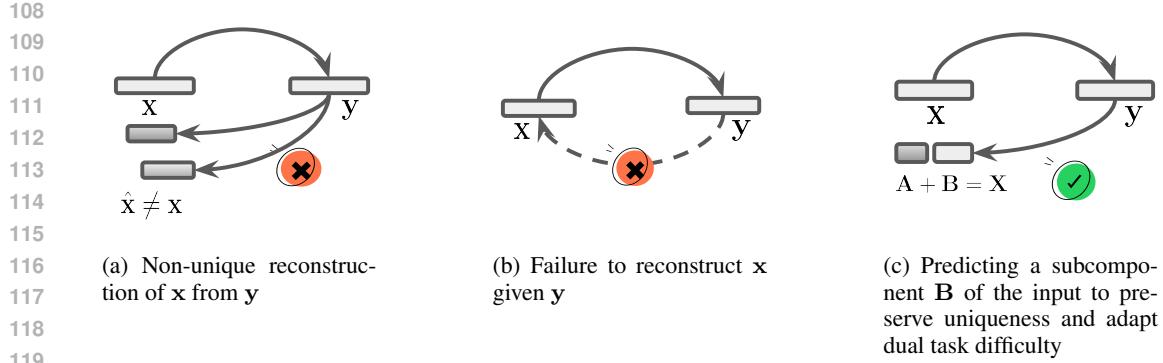
094 Preference optimization steers LLMs’ behavior by assigning scalar rewards to responses based on
095 their quality and higher-reward outputs are preferred and reinforced by the optimization. Formally,
096 given a reward function $r : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ that quantifies output quality, the objective is:

$$097 \quad \max_{\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot | x)} [r(x, y)] \quad (1)$$

099 In practice, $r(x, y)$ is commonly derived from human preferences (RLHF), LLM judgments
100 (RLAIF), or verifiable correctness (RLVR). While numerous effective algorithms (e.g., PPO (Schul-
101 man et al., 2017), REINFORCE++ (Hu et al., 2025a), GRPO (Shao et al., 2024)) have been devel-
102 oped to optimize this objective, their performance ultimately hinges on the quality of the reward
103 signal. Consequently, the key challenge lies in obtaining accurate and scalable rewards, given that
104 existing sources are plagued by high costs, inherent biases, and limited coverage.

105 2.2 DUAL LEARNING

107 Dual learning offers a self-supervision signal by utilizing the task duality. We begin by formalizing
108 the task duality between a *primal* task and its *dual* counterpart.



120 **Figure 1: Challenges in Dual Learning and Solutions via Relaxed Duality Constraints.** Dilem-
 121 mas in applying dual learning — (a) Non-unique reconstruction of x from y breaks the closed-loop;
 122 (b) Failure to reconstruct x from y due to competence asymmetry. Resolutions by relaxing duality
 123 restrictions — (c) Predicting unknown component B to preserve uniqueness and adapt
 124 dual task difficulty

126 **Definition 1.** A primal task is a mapping $\mathcal{T}_p : \mathcal{X} \rightarrow \mathcal{Y}$, and a dual task is a mapping $\mathcal{T}_d : \mathcal{Y} \rightarrow \mathcal{X}$.
 127 The pair $(\mathcal{T}_p, \mathcal{T}_d)$ is said to form a dual pair if they satisfy the consistency principle:

$$129 \quad \forall \mathbf{x} \in \mathcal{X}, \quad d(\mathbf{x}, \mathcal{T}_d(\mathcal{T}_p(\mathbf{x}))) \leq \epsilon_{\mathcal{X}},$$

131 where $d(\cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a domain-specific distance metric, and $\epsilon_{\mathcal{X}} \geq 0$ is a tolerance threshold
 132 that quantifies acceptable reconstruction errors in each space.

133 Leveraging this duality, we can construct a self-supervised reward to quantify the quality of a primal-
 134 task output. Given an input \mathbf{x} and its corresponding output $\mathbf{y} = \mathcal{T}_p(\mathbf{x})$, we could define reward as
 135

$$136 \quad r(\mathbf{x}, \mathbf{y}) \propto \exp(-\lambda \cdot d(\mathbf{x}, \mathcal{T}_d(\mathbf{y}))), \quad (2)$$

137 where $\lambda > 0$ controls the sensitivity of the reward to reconstruction error. High-quality outputs
 138 maximize the expected reward $\mathbb{E}[r(\mathbf{y})]$ by preserving information that is recoverable through the
 139 duality cycle. This principle has been successfully applied in various domains, including machine
 140 translation (He et al., 2016; Zou et al., 2025).

142 3 DUAL LEARNING-BASED PREFERENCE OPTIMIZATION

144 In this section, we propose **Dual Learning-based Preference Optimization (DuPO)**. Its core objec-
 145 tive is to leverage the intrinsic relationships between tasks and their dual counterparts to generate
 146 self-supervised rewards, enabling LLMs to improve performance without relying on expensive hu-
 147 man annotations or complex handcrafted rules.

148 3.1 CHALLENGES IN DUAL LEARNING-BASED OPTIMIZATION

149 While task duality offers a promising self-supervised paradigm, its application to LLM optimization
 150 is non-trivial, as it confronts two critical challenges that disrupt the closed-loop information flow.
 151

152 **Challenge I: Limited Duality in Non-Mutually Implicative Tasks.** The utility of task duality
 153 hinges on \mathcal{T}_p and \mathcal{T}_d being mutually implicative — specifically, the output \mathbf{y} of \mathcal{T}_p contains suffi-
 154 cient information to reconstruct \mathbf{x} via \mathcal{T}_d , and vice versa. This property holds for canonical tasks
 155 like machine translation, where \mathbf{y} (a translation) preserves the semantic content of \mathbf{x} (the source
 156 sentence), enabling \mathcal{T}_d (back-translation) to recover $\hat{\mathbf{x}} \approx \mathbf{x}$.

157 However, most real-world tasks lack this mutual implicativity (Fig. 1a). Consider mathematical
 158 reasoning, where \mathcal{T}_p solves a problem \mathbf{x} (e.g., “A box contains 3 red and 5 blue balls; what is the
 159 total?”) to produce $\mathbf{y} = 8$. Here, \mathbf{y} (the total count) is insufficient to uniquely reconstruct \mathbf{x} via \mathcal{T}_d ,
 160 as 8 could answer infinitely many disparate questions, such as “What is $10 - 2$?” or “What is the
 161 atomic number of Oxygen?”. This underdetermined relationship breaks the duality loop: \mathcal{T}_d cannot
 reliably recover \mathbf{x} from \mathbf{y} , making the self-supervised reward (based on $\hat{\mathbf{x}} \approx \mathbf{x}$) untrustworthy.

162 **Challenge II: Bidirectional Competence Asymmetry.** Even for mutual invertibility tasks, duality
 163 optimization is sensitive to the bidirectional competence of the LLM — the performance gap
 164 between \mathcal{T}_p and \mathcal{T}_d . If \mathcal{T}_p is strong but \mathcal{T}_d is weak, \mathcal{T}_d may produce noisy $\hat{\mathbf{x}}$ that distorts the su-
 165 pervision signal (Fig. 1b). This asymmetry is particularly pronounced in LLMs, where extensive
 166 pretraining creates diverse but uneven capabilities across tasks, even within the same domain.

167 For instance, in machine translation optimization, let \mathbf{x} = “The quick brown fox jumps over the
 168 lazy dog” (English) and \mathbf{y} = “Der schnelle braune Fuchs springt über den faulen Hund” (correct
 169 German translation). A high-quality \mathbf{y} should enable \mathcal{T}_d to back-translate to $\hat{\mathbf{x}} \approx \mathbf{x}$. However, if \mathcal{T}_d
 170 struggles with nuanced vocabulary (e.g., “schnelle” \rightarrow “fast” instead of “quick”), $\hat{\mathbf{x}}$ might be “The
 171 fast brown fox jumps over the lazy dog” — a divergence from \mathbf{x} that erroneously penalizes \mathbf{y} despite
 172 its correctness.

173 Using separate models for \mathcal{T}_p and \mathcal{T}_d , as was common in traditional dual learning, merely sidesteps
 174 the challenge of intrinsic competence asymmetry (Wang et al., 2018). This imbalance still arises
 175 from the distinct natures and complexities of the primal-dual tasks, destabilizing the self-supervised
 176 feedback loop.

178 3.2 GENERALIZED DUALITY FRAMEWORK VIA COMPLEMENTARY TASK

179 To address the two-fold challenges, we propose *generalized duality* that redefines task duality
 180 through complementary dependencies. It transcends traditional duality’s strict input-output rever-
 181 sal requirement by leveraging *partial and stable dependencies* between task components, enabling
 182 robust self-supervised rewarding even for tasks lacking inherent mutual implicativity.

183 **Definition 2.** Let the input space \mathcal{X} of a primal task \mathcal{T}_p be decomposed into two disjoint subspaces:
 184 \mathcal{X}_k (known components) and \mathcal{X}_u (unknown components), such that $\mathcal{X} = \mathcal{X}_k \cup \mathcal{X}_u$. The primal task
 185 \mathcal{T}_p is a mapping $\mathcal{T}_p : \mathcal{X} \rightarrow \mathcal{Y}$ that maps $\mathbf{x} \in \mathcal{X}$ to an output space $\mathbf{y} \in \mathcal{Y}$. Its **complementary dual**
 186 **task** \mathcal{T}_{cd} is a mapping that leverages \mathbf{y} and the known component \mathbf{x}_k to reconstruct the unknown
 187 component $\hat{\mathbf{x}}_u \in \mathcal{X}_u$:

$$188 \mathcal{T}_{cd} : (\mathbf{y}, \mathbf{x}_k) \mapsto \hat{\mathbf{x}}_u.$$

189 The pair $(\mathcal{T}_p, \mathcal{T}_{cd})$ is said to form a *generalized dual pair* if they satisfy the complementary consis-
 190 tency principle:

$$191 \forall \mathbf{x} \in \mathcal{X}, \mathbf{y} = \mathcal{T}_p(\mathbf{x}) : d(\mathbf{x}_u, \mathcal{T}_{cd}(\mathbf{y}, \mathbf{x}_k)) \leq \epsilon,$$

192 where $d(\cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a domain-specific distance metric, and $\epsilon \geq 0$ is a tolerance threshold.

193 Leveraging this generalized duality, we can construct a self-supervised reward to quantify the pref-
 194 erence of a primal-task output analogously to Def. 1. Given an input $\mathbf{x} \in \mathcal{X}$ with decomposition
 195 $x = (x_k, x_u)$ and its corresponding output $y = \mathcal{T}_p(x)$, the reward is defined as

$$196 r(\mathbf{x}, \mathbf{y}) \propto \exp(-\lambda \cdot d(\mathbf{x}_u, \mathcal{T}_{cd}(\mathbf{y}, \mathbf{x}_k))), \quad (3)$$

197 where $\lambda > 0$ controls reward sensitivity.

200 Example 1: Generalized Duality Feedback for a Two-Sum Task: $A + B$

201 The primal task $\mathcal{T}_p : \mathbf{y} \leftarrow \mathbf{x}_u + \mathbf{x}_k$ is to compute the sum of two numbers, with its input and
 202 output as:

203 – The input \mathbf{x} is decomposed as $\mathbf{x} \leftarrow (A, B)$, where $\mathbf{x}_k = A$ (a known number) and $\mathbf{x}_u = B$ (an
 204 unknown number, without loss of generality).

205 – The output \mathbf{y} is the result of sum: $C = A + B$.

206 The complementary dual task $\mathcal{T}_{cd} : \mathbf{x}_u \leftarrow \mathbf{y} - \mathbf{x}_k$ is designed to reconstruct the unknown compo-
 207 nent \mathbf{x}_u , using the primal output \mathbf{y} (i.e. C) and the known \mathbf{x}_k (i.e. A):

$$208 \hat{\mathbf{x}}_u \leftarrow B' = C - A$$

210 Then, we can directly quantify whether B (original unknown) and B' (reconstructed unknown)
 211 are consistent as reward signal:

$$212 r(\mathbf{x}, \mathbf{y}) \propto \exp(-\lambda \cdot \mathbb{I}(B \neq B')).$$

213 Here, $\mathbb{I}(\cdot)$ is an indicator function: it equals 0 if $B = B'$ (consistent) and 1 otherwise (inconsis-
 214 tent). This ensures the reward is maximized when B and B' match, and reduced otherwise.

216 A simple two-sum task is illustrated in Example 1. When generalizing to practical applications,
217 DuPO also splits the input and selects a component of the input (eg., a numerical variable in math
218 problem) to serve as the unknown part x_u and construct the corresponding dual task. To further
219 improve the task duality, the selection could follow principles like *Answerability of the Dual Ques-*
220 *tion* and *Uniqueness of the Correct Completeness* (see Appendix A for details). Moreover, DuPO’s
221 flexibility also allows for task-specific distance metrics $d(\cdot)$. We could employ BLEU scores for
222 multilingual translation, while for mathematical reasoning, we evaluate variable equality, yielding
223 binary rewards, with case studies provided in Appendix D.

224 **Remark 1.** Compared to traditional dual learning, which suffers from strict mutual implicativity
225 (i.e., \mathbf{y} must fully encode \mathbf{x}) and bidirectional competence asymmetry, our generalized duality
226 framework offers three fundamental advantages:

- 227 1. **Overcomes the Invertibility Constraint.** By redesigning the dual objective from recon-
228 structing the full input \mathbf{x} to only a selected unknown component \mathbf{x}_u , our framework fun-
229 damentally bypasses the stringent requirement of task symmetry. This relaxation is the key
230 to unlocking dual learning for tasks that are inherently non-invertible, where the primal
231 output does not contain sufficient information to recover the entire input.
- 232 2. **Mitigates the Competence Asymmetry.** The difficulty of the dual task is significantly re-
233 duced in two ways. First, the known component \mathbf{x}_k acts as a strong contextual anchor,
234 constraining the solution space for reconstruction. Second, we can simply yet effectively
235 select an \mathbf{x}_u that is not only feasibly reconstructible but also act as a faithful reward signal
236 for the primal task’s solution quality (Appendix A). This directly addresses the “weak dual”
237 pitfall and ensures the self-supervised reward is reliable and informative.
- 238 3. **Enables Broad Applicability.** It unlocks dual learning for a broad class of tasks previously
239 considered unsuitable, including complex domains such as mathematical reasoning, code
240 generation, and dialogue systems where input-output relationships are partial.

241 3.3 POLICY OPTIMIZATION WITH COMPLEMENTARY REWARD

242 The core of our Dual Learning-based Preference Optimization (DuPO) framework is to optimize
243 LLMs using duality-derived self-supervised complementary rewards $r(\mathbf{y})$, without external annota-
244 tions. The objective is to maximize the expected reward based on the (complementary) dual task:

$$246 \quad \mathcal{J}(\theta) = \mathbb{E}_{\mathbf{y} \sim \pi_\theta(\mathbf{y}|\mathbf{x})} [r(\mathbf{x}, \mathbf{y})], \quad (4)$$

247 where $\pi_\theta(\mathbf{y}|\mathbf{x})$ denotes the LLM’s policy (parameterized by θ) for generating output \mathbf{y} given in-
248 put $\mathbf{x} = (\mathbf{x}_u, \mathbf{x}_k)$. Notably, DuPO is compatible with various RL algorithms (e.g., PPO, REIN-
249 FORCE++), we adopt GRPO (Shao et al., 2024) in our experiments for its simplicity and efficiency.

250 4 EXPERIMENT

251 We validate the efficacy of DuPO on two representative tasks: multilingual translation and math-
252 ematical reasoning. Below, we detail the experimental setup, datasets, and evaluation metrics for
253 each task, followed by key results.

254 4.1 EXPERIMENT SETUP

255 **Base Model.** We evaluate DuPO on a diverse set of strong and popular base models to demonstrate
256 its effectiveness and robustness. For translation tasks, we employ Seed-X-7B-Instruct (Cheng et al.,
257 2025), one of the strongest open-source translation models. For mathematical reasoning, we select
258 models of varying scales and capabilities, including small-scale yet powerful DeepSeek-R1-Distill-
259 Qwen-1.5B (DeepSeek-AI, 2025) and its larger counterpart DeepSeek-R1-Distill-Qwen-7B, both
260 distilled from the SOTA DeepSeek-R1. We also include Qwen3-4B (Yang et al., 2025), the latest
261 strong small LLM, and the most capable open-source reasoning model, OpenReasoning-Nemotron-
262 7B (Moshkov et al., 2025). These models represent strong and representative baselines within their
263 respective model scales, ensuring comprehensive evaluation. Additionally, we also include some
264 SOTA and impressive ultra-large models like Doubao-1.5/1.6-Thinking (ByteDance Seed Team,
265 2025), Claude-Sonnet4-Thinking, and DeepSeek-R1 (DeepSeek-AI, 2025) for comparison.

266 **Dataset.** For translation tasks, we focus on 28 languages that are aligned with the language cover-
267 age of Seed-X, selecting 1,000 prompts for each language from a multilingual pre-training dataset
268 to create our training prompt set. Additionally, we collect 7,000 parallel data entries across these

Model	BLEU	COMET	BLEURT	Avg.
Qwen3-8B	21.7	84.8	65.8	57.4
Douba0-1.5-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 w/ DuPO (ours)	28.8	87.0	72.6	62.8
w/ DuPO (ours)	30.3	89.1	74.6	64.7

Table 1: **Multilingual Translation Performance Across 756 Translation Directions in 28 Languages.** DuPO significantly improves all metrics and performs comparably to its strong counterparts (DeepSeek models).

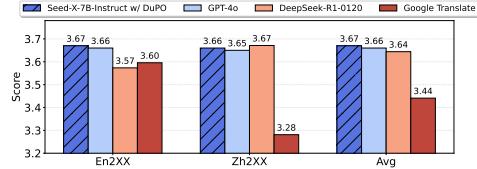


Figure 2: **Human Evaluation Scores (0-4) on the Seed-X-Challenge for 14 Language Directions.** DuPO achieves performance comparable to or even surpassing ultra-large models such as GPT-4o and DeepSeek-R1-0120, while significantly outperforming Google Translate.

specified languages to support our experiments from the dev set of Flores-200 (NLLB Team, 2024). For mathematical reasoning tasks, we utilize questions from a mixture of publicly available mathematics datasets¹. These diverse datasets are widely used to synthesize SFT data or provide oracle labels for RL, training LLMs on complex reasoning tasks.

Benchmarks. For comprehensive evaluation, we include various tasks and benchmarks:

- **Multilingual Translation:** We construct an automatic evaluation set by randomly sampling 50 instances for each of the 756 translation directions (across 28 languages) from the Flores test set². This dataset, comprising 37,800 samples, will be released to facilitate reproducibility. We employ BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and BLEURT (Sellam et al., 2020) as evaluation metrics. For human evaluation, we use the Seed-X-Challenge (Cheng et al., 2025)³, a challenging benchmark to test the boundaries of LLMs’ translation capabilities with diverse linguistic elements across multiple domains. Human experts assess accuracy, fluency, and idiomaticity, scoring translations from Chinese or English to seven languages on a 0-4 scale (higher score denotes better translation).
- **Mathematical Reasoning:** We evaluate our approach on multiple benchmarks, including AMC23 (MAA, 2023), AIME24, AIME25 and HMMT (Balunović et al., 2025), to assess performance in standardized contest environments. For each problem, we sample 32 responses using a temperature of 0.8 and a maximum reasoning budget of 32,000 tokens, then report the average accuracy (Avg@32).

Ultra-large models like DeepSeek-R1 and Douba0-1.6-thinking are accessed via their official APIs. More details about training are provided in Appendix C.

4.2 MAIN RESULTS

4.2.1 DUPO BOOSTS LLM’S PERFORMANCE ON VARIOUS TASKS

DuPO achieves strong performance on diverse tasks, including multilingual translation and mathematical reasoning. On multilingual translation, DuPO elevates the backbone to a SOTA performance level, rivaling and even surpassing significantly ultra-large LLMs. As detailed in Tab. 1, applying DuPO to the Seed-X-7B-Instruct model boosts its performance by 1.5, 2.1, and 2.0 across three automatic evaluation metrics. This performance even surpasses that of current SOTA closed-source ultra-large language models, such as Douba0-1.5-thinking (+2.8) and Qwen3-235B-22B (+1.0), and is on par with the performance of the latest DeepSeek-R1. As shown in Fig. 2, DuPO demonstrates remarkable performance, achieving results comparable to state-of-the-art ultra-large models such as GPT-4o and DeepSeek-R1. Moreover, DuPO substantially outperforms widely-used commercial closed-source systems like Google Translate, showcasing a clear advantage in translation quality as perceived by human evaluators.

On mathematical reasoning, the results in Tab. 2 clearly demonstrate that DuPO yields consistent and significant performance improvements across all models at different scales and baseline reasoning ability. On the most powerful OpenReasoning-Nemotron-7B model, applying DuPO increased

¹More details on math data preparation can be found in Appendix B.

²https://huggingface.co/datasets/openlanguage/flores_plus

³https://github.com/ByteDance-Seed/Seed-X-7B/tree/main/challenge_set

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)

Table 2: **Mathematical Reasoning Performances (%) on Representative Benchmarks.** DuPO significantly improves the performances across models with varying base capabilities, enabling Qwen3-4B to outperform DeepSeek-R1-0120 and OpenReasoning-Nemotron-7B to achieve impressive performance (+7.3).

the average score from 77.1% to 84.4%, achieving impressive performance. This trend of significant gains continues on the mid-sized Qwen3-4B model, which saw its average score boosted by 6.5 points from 67.9% to 74.4%, even surpassing the ultra-large model DeepSeek-R1-0120. The approach remains remarkably effective on DeepSeek’s distilled models as well. Even on DeepSeek-R1-Distill-Qwen-1.5B, the least reasoning capability among the strong baselines, we still achieved a 6.3-point increase in average accuracy. Our framework’s performance is further validated by concrete examples in multilingual translation and mathematical reasoning (case studies in Appendix D).

4.2.2 DUPO SCALES TO VARIOUS BACKBONES EFFECTIVELY

To validate the robustness and generalization of our proposed DuPO framework, we extend our evaluation to the LLaMA architectural family. Our experiments are conducted on two LLaMA architectural models: LLaMA-3.1-8B (Grattafiori et al., 2024) and OctoThinker-8B-Hybrid-Base (Wang et al., 2025), the latter of which has undergone middle training on mathematical reasoning knowledge. Considering the significant difference in model ability, we select two benchmarks of moderate difficulty, AMC23 (MAA, 2023) and MATH500 (Hendrycks et al., 2021). For a fair comparison, all models are finetuned using identical training data and settings. Results are listed in Tab. 3.

As seen, DuPO’s effectiveness is not tied to a specific model architecture; it serves as a robust and generalizable enhancement, delivering significant improvements to diverse backbones regardless of their initial reasoning proficiency. DuPO lifts the average score of LLaMA-3.1-8B to 32.1%, a +24.0 percentage-point gain over the vanilla model, and surpasses SimpleRL-Zoo (Zeng et al., 2025) (which relies on oracle-labeled answers during training) by 13.1%. When applied to the OctoThinker-8B-Hybrid-Base (Wang et al., 2025), our DuPO approach yields even more impressive performance improvements of +50.0 on AMC23 and +27.4 on MATH500, achieving an average performance of 62.5.

4.2.3 DUPO INCENTIVIZES REASONING CAPABILITY ON BASE MODEL

Further validating DuPO’s versatility, we demonstrate its effectiveness even when applied directly to a base model without a preliminary supervised fine-tuning (SFT) phase for activation of reasoning behavior. We track the learning dynamics by simultaneously collecting the primal task accuracy (“Forward Acc”) and dual task accuracy (“Backward Acc”) on the training set and its generalization performance on three unseen test set.

As shown in Fig. 3, DuPO provides a stable and effective pathway to awaken and generalize the latent reasoning abilities of a base model, validating its utility as a powerful training methodology. Specifically, the training dynamics reveal a clear and substantial improvement on the primal task, with the “Forward Acc” soaring from a nascent 15.2% to 56.5%. Our automated unknown-component selection strategy (based on Qwen3-4B-Instruct) adjusts dual-question difficulty, yield-

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

378
379
380
381
382
383
384
385
386 **Table 3: Performances (%) of DuPO on Different Backbone Models.** DuPO even surpasses SimpleRL-Zoo, which utilizes labeled answers as reward. DuPO’s potential is further exemplified by OctoThinker, which underwent additional middle training.
391

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-Qwen-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)

400
401 **Table 4: Inference-Time Scaling on Mathematical Reasoning Using DuPO Rewarding (Backward Acc) for Reranking.** Our method improves the performance of policy models with varying base ability, without requiring additional training.
402
403
404

405 ing relatively high initial accuracy (52.6%) and effective rewards. As the model improves during
406 training, it naturally solves more dual questions, thereby unlocking richer reward signals. This up-
407 ward trajectory provides direct evidence that the reward signal derived from our dual-task serves
408 as an effective guide for enhancing the model’s reasoning. More importantly, this acquired skill
409 demonstrates robust generalization. Performance on the test set AMC23 leaped from 20% to 70%,
410 with similarly significant gains observed on the AIME24 and AIME25 datasets.
411

4.2.4 DUPO SCALES REASONING DURING INFERENCE WITHOUT TRAINING

412 Beyond serving as a reward signal for RL training, the DuPO mechanism can be naturally applied as
413 a training-free, inference-time reranking strategy to improve the reasoning capabilities of any LLM.
414 The process unfolds in three stages: 1) Similar to the rollout stage during RL process, we could
415 prompt any given policy model to generate diverse reasoning trajectories. 2) For each candidate
416 trajectory, we use its final answer to ask the policy model to solve the corresponding dual question
417 automatically constructed without accessing labeled answer. We could apply more computation by
418 performing K ($K = 8$ in our experiments) sampling runs on each dual question for a more reliable
419 reward estimate, a practice distinct from RL training. 3) Finally, for each test set question, we select
420 the trajectory with the highest backward accuracy on its dual questions as the final output.
421

422 As presented in Tab. 4, the experimental results demonstrate that DuPO provides accurate reward
423 signals, effectively guiding models towards correct reasoning, serving as an efficient approach for
424 scaling reasoning capabilities even without training. On the two challenging AIME benchmarks,
425 applying DuPO as a reranking method improves the average performance of Qwen3-4B by 9.3
426 points, elevating its accuracy from 68.4% to 77.7% without any additional training. Notably,
427 the DuPO-enhanced Qwen3-4B surpasses DeepSeek-R1 and Claude-Sonnet4-Thinking (77.7% vs.
428 74.9%/76.3% on average). The impact on DeepSeek-R1-Distill-Qwen-1.5B is even more pro-
429 nounced, with an 18.7 point increase (20.0% to 38.7%).
430

4.3 EFFECTS OF TASK DUALITY

431 To verify the efficacy of our component selection strategy, we conduct an ablation study by removing
432 it and training on the entire unfiltered dataset. The results, illustrated in Figure 4, reveal a clear

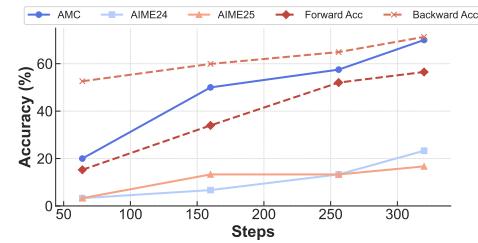


Figure 3: **Training Progress of DuPO on Qwen3-4B-Base.** The performance consistently improves on the primal/dual task (Forward/Backward Acc) and the benchmarks.
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430

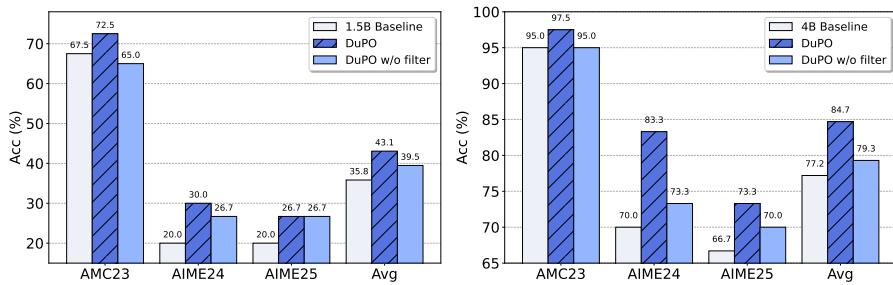


Figure 4: **Performance Ablation of DeepSeek-R1-Distill-Qwen-1.5B/Qwen-3-4B on Mathematical Reasoning.** Our unknown component selection strategy reduces training noise and improves these models’ performance across three benchmarks.

trend: removing the selection strategy results in a significant performance degradation across most benchmarks. This is further evidenced by an average accuracy drop of 3.6 and 5.4 percentage points for the 1.5B and 4B models, respectively. The result validates our strategy’s effectiveness, confirming that by improving task duality, it provides a cleaner reward signal that is crucial for achieving superior performance.

5 RELATED WORK

5.1 PREFERENCE OPTIMIZATION FOR LLMs

Preference optimization is pivotal for aligning LLMs with desired behaviors, with current research dominated by three paradigms reliant on external supervision. RLHF (Ouyang et al., 2022) aligns models with human preferences, but is fundamentally hampered by the cost and inconsistency of human annotation (Lee et al., 2023; Zhang et al., 2024). As a cost-effective alternative, LLM-as-a-Judge (Zheng et al., 2023; Lee et al., 2023) utilizes a powerful LLM for evaluation. The reliability is perpetually undermined by the judge model’s own performance limitations and intrinsic biases, such as sensitivity to response ordering or stylistic artifacts (Wang et al., 2024; Gudibande et al., 2023; Li et al., 2024a). In parallel, RLVR has shown success in domains like mathematics by using ground-truth outcomes as reward signals (DeepSeek-AI, 2025; Team, 2025; Yang et al., 2025). However, this paradigm’s continued reliance on labeled answers as external supervision and it’s ill-suited for free-form tasks, such as multilingual translation, that inherently lack a single, definitive ground truth. Recent self-supervised paradigms explore alternative reward sources. Some leverage consistency checks between problem paraphrases (Zhang et al., 2025), while others utilize self-play to generate a curriculum (Huang et al., 2025), which requires managing a multi-agent adversarial process. **RLT** (Cetin et al., 2025) focuses on training a teacher model to generate more effective distillation data, by rewarding the teacher’s explanations based on the student’s performance. DuPO differs by sourcing its reward from the intrinsic, dual structure of a task itself and provides a reliable and self-contained verification signal, sidestepping dependencies on auxiliary data generation or the complexities of adversarial dynamics.

5.2 DUAL LEARNING

Dual learning enhances model performance by leveraging intrinsic task symmetry, where a primal task and its complementary dual task mutually provide supervision. He et al. (2016) first introduced dual learning for machine translation, which uses bidirectional tasks (e.g., En→Zh and Zh→En) to generate pseudo-labels via back-translation (Sennrich et al., 2015), reducing reliance on parallel corpora—a breakthrough for low-resource language pairs. Building on this foundation, the paradigm has proven highly versatile—spanning multi-modal (Yi et al., 2017; Zhu et al., 2017; Ren et al., 2020) and knowledge reasoning (Dognin et al., 2020), and extending to reinforcement learning (Luo et al., 2019; Bahng et al., 2025). In modern LLMs, it further refines output quality and enforces semantic consistency (Zou et al., 2025; Chen et al., 2024). However, the reliance on strict task duality, requiring mutually invertible tasks—precludes its application to open-ended or creative domains. In this work, we generalize the dual learning paradigm, moving beyond this rigid invertibility constraint.

486 6 CONCLUSION 487

488 We introduce DuPO, a dual learning-based preference optimization framework that eliminates the
489 need for costly human annotations and handcrafted rewards in LLM training. Its core innovation
490 *generalized duality*, generates self-supervised feedback by decomposing and reconstructing
491 input spaces, addressing critical limitations of traditional dual learning and preference optimization
492 paradigms. DuPO’s effectiveness is validated across diverse tasks: in mathematical reasoning, it
493 boosts average accuracy by 6.4 percent points across models from 1.5B to 7B, while in multilingual
494 translation, it elevates a 7B model to rival larger SOTA LLMs with COMET score gains of up to 2.1
495 across 756 translation directions. Furthermore, as a training-free reranker, DuPO enables smaller
496 models to outperform significantly larger LLMs by up to 9.3 points. This model-agnostic and task-
497 versatile design positions DuPO as a scalable, annotation-efficient solution for more autonomous,
498 adaptable, and cost-effective LLM optimization.

499 REPRODUCIBILITY STATEMENT

500 To ensure the full reproducibility of our work, we provide a comprehensive description of our
501 methodology and experimental setup. We detail the step-by-step application of our method in Ap-
502 pendix A and the pipeline for training data construction in Appendix B. For training and evaluation,
503 Section 4.1 specifies the large language models and datasets utilized, while Appendix C lists the
504 corresponding hyperparameters and configurations used. Furthermore, illustrative case studies are
505 presented in Section D to offer qualitative insights into our method’s application. To further promote
506 reproducibility and facilitate future research, we will also open-source our code.

507 REFERENCES

509 Mistral AI. Mistral large. <https://mistral.ai/news/mistral-large-2407>, 2024.

510

511 Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait
512 Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-Math: A large-scale,
513 high-quality math dataset for reinforcement learning in language models, 2025. URL <https://arxiv.org/abs/2502.17387>.

515 Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learn-
516 ing image-text alignment without human preferences. *arXiv preprint arXiv:2506.02095*, 2025.

517

518 Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones,
519 Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Ols-
520 son, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-
521 Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse,
522 Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mer-
523 cado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna
524 Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Con-
525 erly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario
526 Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai:
527 Harmlessness from ai feedback, 2022. URL <https://arxiv.org/abs/2212.08073>.

528 Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena:
529 Evaluating llms on uncontaminated math competitions, 2025.

530

531 ByteDance Seed Team. Seed1.5-Thinking: Advancing superb reasoning models with reinforcement
532 learning, 2025. URL <https://arxiv.org/abs/2504.13914>.

533 Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine
534 translation research. In Diana McCarthy and Shuly Wintner (eds.), *11th Conference of the Eu-
535 ropean Chapter of the Association for Computational Linguistics*, pp. 249–256, Trento, Italy,
536 April 2006. Association for Computational Linguistics. URL <https://aclanthology.org/E06-1032/>.

538

539 Edoardo Cetin, Tianyu Zhao, and Yujin Tang. Reinforcement learning teachers of test time scaling,
2025. URL <https://arxiv.org/abs/2506.08388>.

540 Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao,
541 and Min Zhang. DUAL-REFLECT: Enhancing large language models for reflective translation
542 through dual learning feedback mechanisms, 2024. URL <https://arxiv.org/abs/2406.07232>.

543

544 Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catan-
545 zaro, and Wei Ping. AceReason-Nemotron: Advancing math and code reasoning through rein-
546 forcement learning. *arXiv preprint arXiv:2505.16400*, 2025.

547

548 Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu,
549 Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie
550 She, Lu Xu, Nuo Xu, Sen Yang, Runsheng Yu, Yiming Yu, Liehao Zou, Hang Li, Lu Lu, Yuxuan
551 Wang, and Yonghui Wu. Seed-X: Building strong multilingual translation llm with 7b parameters,
552 2025. URL <https://arxiv.org/abs/2507.13618>.

553 DeepMind. Gemini 2.5. <https://deepmind.google/technologies/gemini/>, 2025.
554 Accessed: 2025-04-18.

555 DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning,
556 2025.

557

558 Pierre L. Dognin, Igor Melnyk, Inkit Padhi, Cícero Nogueira dos Santos, and Payel Das. Dualtkb:
559 A dual learning bridge between text and knowledge base. In Bonnie Webber, Trevor Cohn, Yulan
560 He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural
561 Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 8605–8616. Associa-
562 tion for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.694. URL
563 <https://doi.org/10.18653/v1/2020.emnlp-main.694>.

564 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
565 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd
566 of models. *ArXiv preprint*, abs/2407.21783, 2024. URL <https://arxiv.org/abs/2407.21783>.

567

568 Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey
569 Levine, and Dawn Song. The false promise of imitating proprietary llms. *CoRR*, abs/2305.15717,
570 2023. doi: 10.48550/ARXIV.2305.15717. URL <https://doi.org/10.48550/arXiv.2305.15717>.

571

572 Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual
573 learning for machine translation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg,
574 Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Sys-
575 tems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10,
576 2016, Barcelona, Spain*, pp. 820–828, 2016. URL <https://proceedings.neurips.cc/paper/2016/hash/5b69b9cb83065d403869739ae7f0995e-Abstract.html>.

577

578 Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang
579 Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Yang Liu,
580 and Yahui Zhou. Skywork open reasoner series, 2025. Notion Blog.

581

582 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,
583 and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv
584 preprint*, abs/2103.03874, 2021. URL <https://arxiv.org/abs/2103.03874>.

585

586 Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: An efficient rlhf algorithm
587 with robustness to both prompt and reward models, 2025a. URL <https://arxiv.org/abs/2501.03262>.

588

589 Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang, and Heung-Yeung Shum.
590 Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base
591 model, 2025b. URL <https://arxiv.org/abs/2503.24290>.

592

593 Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin
Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data, 2025. URL
<https://arxiv.org/abs/2508.05004>.

594 Yunjie Ji, Sitong Zhao, Xiaoyu Tian, Haotian Wang, Shuaiting Chen, Yiping Peng, Han Zhao, and
595 Xiangang Li. How difficulty-aware staged reinforcement learning enhances llms' reasoning capa-
596 bilities: A preliminary experimental study, 2025. URL <https://arxiv.org/abs/2504.00829>.

597

598 Ruipeng Jia, Yunyi Yang, Yongbo Gai, Kai Luo, Shihao Huang, Jianhe Lin, Xiaoxi Jiang, and
599 Guanjun Jiang. Writing-zero: Bridge the gap between non-verifiable tasks and verifiable rewards,
600 2025. URL <https://arxiv.org/abs/2506.00103>.

601

602 Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Vic-
603 tor Carbune, and Abhinav Rastogi. RLAIF vs. RLHF: Scaling reinforcement learning from hu-
604 man feedback with ai feedback. In *International Conference on Machine Learning*, 2023. URL
605 <https://api.semanticscholar.org/CorpusID:261493811>.

606

607 Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita
608 Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From
609 generation to judgment: Opportunities and challenges of llm-as-a-judge. *CorR*, abs/2411.16594,
610 2024a. doi: 10.48550/ARXIV.2411.16594. URL <https://doi.org/10.48550/arXiv.2411.16594>.

611

612 Jiahuan Li, Shanbo Cheng, Shujian Huang, and Jiajun Chen. Mt-patcher: Selective and extendable
613 knowledge distillation from large language models for machine translation, 2024b. URL <https://arxiv.org/abs/2403.09522>.

614

615 Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. Eliciting the translation abil-
616 ity of large language models via multilingual finetuning with translation instructions. *Transactions*
617 *of the Association for Computational Linguistics*, 12:576–592, 2024c. doi: 10.1162/tacl_a_00655.
618 URL <https://aclanthology.org/2024.tacl-1.32/>.

619

620 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
621 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *ArXiv preprint*,
622 abs/2412.19437, 2024. URL <https://arxiv.org/abs/2412.19437>.

623

624 Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. A
625 dual reinforcement learning framework for unsupervised text style transfer. In Sarit Kraus
626 (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelli-
627 gence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 5116–5122. ijcai.org, 2019. doi:
628 10.24963/IJCAI.2019/711. URL <https://doi.org/10.24963/ijcai.2019/711>.

629

630 MAA. Amc 2023 problems, 2023. URL https://artofproblemsolving.com/wiki/index.php/2023_AMC_12A_Problems. Accessed: 2025-05-11.

631

632 Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt
633 Schifferer, Wei Du, and Igor Gitman. AIMO-2 Winning Solution: Building state-of-the-art math-
634 ematical reasoning models with openmathreasoning dataset, 2025. URL <https://arxiv.org/abs/2504.16891>.

635

636 NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846,
637 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07335-x. URL <https://doi.org/10.1038/s41586-024-07335-x>.

638

639 OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2023. Accessed:
640 2025-04-18.

641

642 Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong
643 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kel-
644 ton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike,
645 and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.
646 URL <https://arxiv.org/abs/2203.02155>.

647

648 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic
649 evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.),
650 *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.

648 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguis-
649 tics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040/>.
650

651 Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for mt
652 evaluation, 2020. URL <https://arxiv.org/abs/2009.09025>.
653

654 Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Almost unsupervised text to
655 speech and automatic speech recognition, 2020. URL <https://arxiv.org/abs/1905.06791>.
656

657 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
658 optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL <http://arxiv.org/abs/1707.06347>.
659

660 Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text
661 generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings*
662 of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881–7892,
663 Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.
664 704. URL [https://aclanthology.org/2020.acl-main.704/](https://aclanthology.org/2020.acl-main.704).
665

666 Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation mod-
667 els with monolingual data. *CoRR*, abs/1511.06709, 2015. URL <http://arxiv.org/abs/1511.06709>.
668

669 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
670 Mingchuan Zhang, YK Li, Y Wu, et al. DeepseekMath: Pushing the limits of mathematical
671 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
672

673 Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. *ArXiv*, abs/2501.12599, 2025.
674 URL <https://api.semanticscholar.org/CorpusID:275789974>.
675

676 Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng
677 Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In
678 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meet-
679 ing of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450,
680 Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/
681 2024.acl-long.511. URL <https://aclanthology.org/2024.acl-long.511/>.
682

683 Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. Dual transfer
684 learning for neural machine translation with marginal distribution regularization. In Sheila A.
685 McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on
686 Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-
687 18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18),
688 New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5553–5560. AAAI Press, 2018. doi:
689 10.1609/AAAI.V32I1.11999. URL <https://doi.org/10.1609/aaai.v32i1.11999>.
690

691 Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes
692 reinforcement learning scaling, 2025. URL <https://arxiv.org/abs/2506.20512>.
693

694 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
695 Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,
696 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin
697 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,
698 Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,
699 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *ArXiv preprint*,
700 abs/2412.15115, 2024a. URL <https://arxiv.org/abs/2412.15115>.
701

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu,
Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu,
Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math Technical Report: Toward mathematical ex-
pert model via self-improvement, 2024b. URL <https://arxiv.org/abs/2409.12122>.

702 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
703 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,
704 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin
705 Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,
706 Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui
707 Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang
708 Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger
709 Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan
710 Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

711 Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning
712 for image-to-image translation. In *IEEE International Conference on Computer Vision, ICCV*
713 *2017, Venice, Italy, October 22-29, 2017*, pp. 2868–2876. IEEE Computer Society, 2017. doi:
714 10.1109/ICCV.2017.310. URL <https://doi.org/10.1109/ICCV.2017.310>.

715 Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong
716 Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi
717 Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi
718 Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-
719 Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An
720 open-source llm reinforcement learning system at scale, 2025.

721 Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zeyun Ma, and Junxian He. Simplerl-
722 zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025.
723 URL <https://arxiv.org/abs/2503.18892>.

724 Michael JQ Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol
725 Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree
726 and do models know?, 2024. URL <https://arxiv.org/abs/2410.14632>.

727 Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao
728 Yao, and Bo Han. Co-reward: Self-supervised reinforcement learning for large language
729 model reasoning via contrastive agreement, 2025. URL <https://arxiv.org/abs/2508.00410>.

730 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao
731 Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez,
732 and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh,
733 Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.),
734 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural In-*
735 *formation Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10*
736 *- 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

737 Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation
738 using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference*
739 *on computer vision*, pp. 2223–2232, 2017.

740 Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen,
741 and Lei Li. Multilingual machine translation with large language models: Empirical results and
742 analysis, 2024. URL <https://arxiv.org/abs/2304.04675>.

743 Wei Zou, Sen Yang, Yu Bao, Shujian Huang, Jiajun Chen, and Shanbo Cheng. TRANS-ZERO:
744 Self-play incentivizes large language models for multilingual translation without parallel data.
745 In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.),
746 *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12337–12347, Vi-
747 enna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-
748 5. doi: 10.18653/v1/2025.findings-acl.637. URL <https://aclanthology.org/2025.findings-acl.637>.

749

756 **A CONSTRUCTION OF DUAL QUESTION FOR MATH REASONING**
757

758 We propose a simple approach for construction of dual question for mathematical reasoning. The
759 algorithm operates on mathematical expressions and performs the following key steps:

760 1. **Pattern Recognition and Exclusion:** The algorithm first identifies numerical candidates
761 within the expression while excluding numbers in specific contexts: subscripts (x_1, x_2), in-
762 equality constraints ($x \leq 5$), common exponential bases ($2^n, 10^k$), and function arguments
763 ($f(3)$).

764 2. **Variable Generation and Replacement:** For each valid numerical candidate, the system
765 generates a unique variable identifier of the form $\text{Variable}_{\{str\}}$ where str is a randomly
766 generated lowercase string. The original number is then substituted with this variable.

767 3. **Question Generation of Dual Task:** Using the transformed expression and the original
768 answer, the algorithm constructs inverse problems following templates such as: “Given
769 that the correct answer is $\{answer\}$, determine the value of $\{variable\}$.”

771 This methodology enables systematic generation of problem variants while preserving mathematical
772 validity and semantic coherence. From a single primal question, multiple dual questions can be
773 derived. To ensure that these dual questions robustly satisfy the properties of duality, we filter the
774 candidates using the following two principles:

775 1. **Answerability of the Dual Question:** For the set of sampled answers collected for a given
776 primal question, at least one answer must be capable of correctly solving the corresponding
777 dual question.

778 2. **Uniqueness of the Correct Completeness:** Among the same set of sampled answers, at
779 most one should correctly answer the dual question.

780 Taken together, these two principles ensure that for any selected dual question, there is one and only
781 one correct answer within the pool of candidate solutions for the primal task. This establishes the
782 one-to-one correspondence necessary for generating a reliable self-supervised reward signal.

784 **B MATH RL DATASET PREPARATION**
785

786 Our dataset preparation process began with the collection of 1,815,942 prompts from various pub-
787 licly available datasets (Chen et al., 2025; Albalak et al., 2025; He et al., 2025; Ji et al., 2025).
788 After deduplication, we obtained 318,649 primal questions and generated 1,059,671 dual questions
789 through our designed steps as discussed above. After that, we employed Qwen2.5-7B-Instruct (Yang
790 et al., 2024a) to sample 32 candidate answers for each primal question and then prompted it to an-
791 swer the corresponding dual question based on these candidates. Subsequently, we rigorously fil-
792 tered out all dual questions that failed to meet our predefined principles above. We repeated this
793 sampling and filtering process with Qwen3-4B (Yang et al., 2025), this time with 8 candidate an-
794 swers per question. The resulting collection of high-quality, diverse mathematical questions along
795 with corresponding dual questions formed our final RL training set, providing a robust foundation
796 for our reinforcement learning process in the mathematical domain.

797 **C EXPERIMENT DETAILS**

798 We present more details about our training as follows. For the training process, we use a train batch
799 size of 512, mini batch size of 32, sampling temperature of 1.0, and 16 rollouts per prompt. **We**
800 **employ the AdamW optimizer** with a learning rate of 1e-6 and weight decay of 0.1 (following the
801 default configuration in verl), with gradient clipping set to 1.0. For translation tasks, we set the
802 maximum input length to 2,048 tokens and output length to 4,096 tokens. For mathematical tasks,
803 we use the same input length but extend the maximum output length to 30,000 tokens.

804 **D CASE STUDY**

806 To illustrate the efficacy of our DuPO approach, we present two representative scenarios in Tab. 5
807 that demonstrate how DuPO provides a reliable reward signal across diverse domains.

808 **Scenario 1: Mathematical Reasoning Validation.** In mathematical reasoning, DuPO derives dual
809 task questions from the primal task question where key numerical parameters are replaced with

810	Scenario 1: DuPO on Mathematical Reasoning	
811	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)
812	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 answer , what must V_{sk} have been?
813	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 answer , what's the corresponding V_{rj} ?
814	Candidates	Answer: 468 Backward Accuracy: 69.1% Answer: 108 Backward Accuracy: 0% Answer: 312 Backward Accuracy: 0%
815	Scenario 2: DuPO on Machine Translation (MT)	
816	Primal Task	Translate to Chinese: As knowledge of Greek declined, the West found itself cut off from its Greek philosophical and scientific roots.
817	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。
818	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)
819	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)
820	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)
821	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)

838 **Table 5: Case Studies of DuPO on Mathematical Reasoning and Machine Translation.** DuPO
 839 validates each candidate's quality through a corresponding dual task, reliably identifies **the superior**
 840 **solution** over **inferior ones**.

841
 842
 843 variables, and the model tries to work backwards conditioned on candidate answers. When given
 844 a geometry problem about triangle properties, three candidate answers are sampled: 468, 108, and
 845 312. DuPO automatically derives two dual questions by replacing the circumradius (13) and inradius
 846 (6) with variables, asking the model to deduce these values from the proposed answer. The candidate
 847 answer 468 achieves 69.1% accuracy on dual task, while the incorrect answers (108 and 312) totally
 848 fail to answer the dual task.

849 **Scenario 2: Machine Translation Quality Assessment.** For translation tasks, DuPO leverages re-
 850 verse direction translation as the dual task to evaluate translation quality. Given an English sentence
 851 about Greek philosophical decline, two Chinese translation candidates are generated and subse-
 852 quently back-translated to English. The first translation achieves a BLEU score of 45.85 in the
 853 forward direction and 82.07 in the back-translation, demonstrating semantic preservation and trans-
 854 lation fidelity. In contrast, the second candidate shows degraded performance with BLEU scores of
 855 28.65 and 16.11, respectively, indicating semantic drift and poor translation quality.

856 These case studies validate DuPO's core hypothesis: high-quality solutions maintain consistent in-
 857 formation across dual task formulations, while inferior solutions exhibit significant degradation.
 858 This dual validation mechanism provides a robust framework for automatic quality assessment with-
 859 out requiring ground truth labels.

860 861 **E PERFORMANCE GAP ANALYSIS BETWEEN DUPO AND ORACLE**

862
 863 Further analysis reveals that DuPO achieves near-optimal performance, effectively approaching the
 results obtainable with ground-truth supervision. To quantify this gap, we construct two oracle

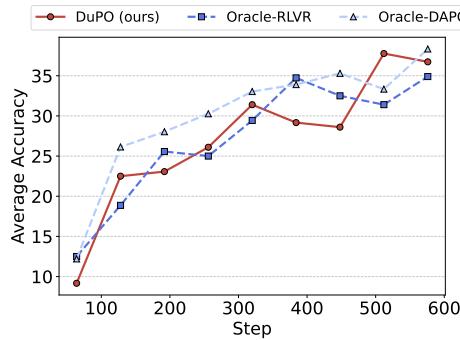


Figure 5: **Training curves of DuPO vs. Oracle-RLVR/Oracle-DAPO.** DuPO closely tracks the oracle baseline that uses ground-truth answer for verifiable reward throughout training.

baselines, **Oracle-RLVR** and **Oracle-DAPO**, where both utilize ground-truth answers to verify rollouts as reward signals. Specifically, Oracle-RLVR utilizes the exact same dataset as DuPO, serving as a direct baseline. In contrast, Oracle-DAPO employs the DAPO Yu et al. (2025) dataset, which is a representative high-quality annotated dataset from the open-source community. These oracles indicate the performance upper bound of the RLVR paradigm given accurate supervision. For efficiency, we conducted experiments using Qwen3-4B-Base with 8192 maximum output length.

As shown in Figure 5, DuPO closely tracks Oracle-RLVR throughout the training process. By step 600, both methods reach nearly identical performance levels at approximately 35% accuracy. This close alignment indicates that our self-verification reward signals are accurate and effective. Furthermore, we observe that the performance of Oracle-DAPO is comparable to that of both DuPO and Oracle-RLVR. This similarity suggests that our x_u selection strategy has a minimal impact on the data distribution, thereby demonstrating the robustness of DuPO.

F LIMITATION DISCUSSION

DuPO has demonstrated its effectiveness on mathematical reasoning (representing complex reasoning tasks) and machine translation (representing constrained generation tasks), providing a novel pathway for self-supervised training in LLMs. However, fully open-ended tasks remain inherently challenging. Tasks such as creative writing, where outputs are highly unconstrained and evaluation criteria are subjective, require further exploration to effectively extract the unknown component and construct meaningful dual problems. Intuitively, taking creative writing as an example, specific constraints in user instruction (e.g., theme, style, tone) might be extracted as the unknown component x_u for reconstruction. However, the practical implementation and validation of such mechanisms in these open-ended domains constitute important directions for future work.

G THE USE OF LARGE LANGUAGE MODELS

We utilized a Large Language Model (LLM) as an assistive tool to proofread the manuscript for grammatical errors and to provide suggestions for improving clarity and flow. All suggestions were manually reviewed by the authors, who made the final decisions on all textual modifications to enhance the paper’s readability.