WHEN RIGHT MEETS WRONG: BILATERAL CONTEXT CONDITIONING WITH REWARD-CONFIDENCE CORRECTION FOR GRPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Relative Policy Optimization (GRPO) has emerged as an effective method for training reasoning models. Although GRPO computes advantages from the group mean, it treats each output as an independent sample during optimization and leaves unused the contrast between correct and incorrect solutions within the same group, where successful reasoning traces could be compared explicitly against failed ones. To exploit this contrast, we present a contrastive reformulation of GRPO, showing that the GRPO objective implicitly maximizes the margin between the policy ratios of correct and incorrect samples. Building on this insight, we propose Bilateral Context Conditioning (BICC), a mechanism that allows the model to cross-reference successful and failed reasoning traces during the optimization, enabling a direct information flow across samples. We further introduce Reward-Confidence Correction (RCC) to stabilize training by dynamically adjusting the advantage baseline in GRPO using reward-confidence covariance derived from the first-order approximation of the variance-minimizing estimator. Both mechanisms operate on the sampled group alone and apply to GRPO and its variants. Experiments with Qwen3-4B and Phi-4-mini on Math500, AMC 2023, and AIME 2024/2025 show that BICC improves GRPO and five of its variants in 54 of 56 comparisons, by 0.3 to 1.9 points with larger gains on the weaker base model, while RCC brings gradient variance 31 to 37% below GRPO. The gains persist at 14B scale and on Llama-3.1-8B, and training on mathematics transfers zero-shot to logic and code. The implementation is available at BiCC Codebase.

1 INTRODUCTION

Large language models (LLMs) have improved their reasoning through Reinforcement Learning with Verifiable Rewards (RLVR), where models learn from outcome-based feedback such as binary correctness signals on mathematical problems (Lightman et al., 2023; Shao et al., 2024; Guo et al., 2025). Group Relative Policy Optimization (GRPO) has become the prevalent optimizer: in place of the critic of Proximal Policy Optimization (Schulman et al., 2017), it samples several solutions per query and estimates advantages from their relative performance (Guo et al., 2025), a design adopted by several state-of-the-art reasoning models (Mroueh et al., 2025; Yang et al., 2025).

Vanilla GRPO leaves a signal inside each group unused: in a group of eight solutions partitioned into correct and incorrect subsets, correct solutions tend to share successful strategies while incorrect ones reveal common failure modes (Uesato et al., 2022; Tyen et al., 2024). Yet GRPO evaluates each solution in isolation against the group mean and remains blind to the rest of the group beyond the advantage (Yue et al., 2025), so the contrast between successful and failed paths on the same query goes unused. We formalize the observation by reformulating the GRPO objective into a contrastive form, which reveals that GRPO implicitly maximizes the margin between the average policy ratios of correct and incorrect samples and makes the partition structure visible. Building on the insight, we propose Bilateral Context Conditioning (BICC), which enables explicit cross-partition information flow. The model observes incorrect attempts when evaluating correct ones, and vice versa, and the opposite-partition samples serve as privileged context available only during training, in the sense of Learning Using Privileged Information (LUPI; Pechyony & Vapnik, 2010), so the contrast leaves the inference cost unchanged.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

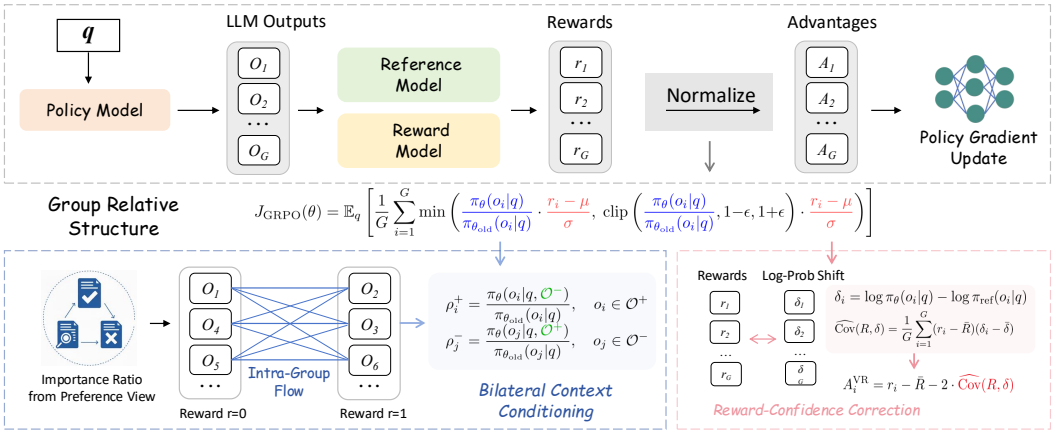


Figure 1: The figure contrasts the standard GRPO pipeline with the two proposed components. The upper row is the standard GRPO pipeline, kept as the reference: the policy generates G outputs for query q, a verifier scores them, and the normalized advantages drive the policy gradient. The lower row marks the two components that change. BICC partitions the outputs by reward, builds for each sample an augmented context from the opposite partition within a fixed token budget, matching the smaller partition to several members of the larger one, and evaluates the sample under that context in the numerator of the importance ratio. RCC computes the covariance between rewards and log-probability shifts within the group and subtracts it from the advantage.

To stabilize training under bilateral conditioning, we further introduce Reward-Confidence Correction (RCC). First-order approximation of the variance-minimizing baseline under importance sampling yields a covariance correction that accounts for the correlation between the confidence of the model and the reward and reduces gradient variance from quantities of the sampled group alone.

The two mechanisms act on complementary components: BICC modifies the policy ratio and RCC the advantage. We evaluate on Qwen3-4B and Phi-4-mini across Math500, AMC 2023, AIME 2024, and AIME 2025. BICC improves 54 of the 56 comparisons by 0.3 to 1.9 points, with larger gains on weaker base models, and RCC brings gradient variance 31 to 37% below GRPO.

❶ We reformulate GRPO contrastively and show that its objective maximizes the margin between the policy ratios of correct and incorrect samples of a group.
❷ We propose BICC, which lets correct and incorrect attempts inform each other through cross-partition context, and RCC, a first-order approximation of the optimal baseline that reduces gradient variance through the reward-confidence covariance.
❸ We validate both mechanisms across GRPO and five of its variants on two base models, with gains on competition-level mathematical reasoning benchmarks.

2 Related Work

Reinforcement learning for LLM reasoning moved from RLHF (Christiano et al., 2017; Ouyang et al., 2022) to training on verifiable answers (Jaech et al., 2024; Guo et al., 2025), and GRPO (Shao et al., 2024) removes the critic by normalizing rewards within a group. The variants of GRPO change the reference policy (Liu et al., 2025), the clipping bounds (Yu et al., 2025), or the probability aggregation (Zhao et al., 2025; Zheng et al., 2025), and these variants evaluate each sample in isolation from its group. Offline preference methods such as DPO (Rafailov et al., 2023) and IPO (Saeidi et al., 2025) contrast pre-collected pairs through a closed-form likelihood that dispenses with importance ratios, advantages, and baselines, a separate paradigm from on-policy group optimization. Baselines are the classical tool for variance reduction (Williams, 1992); under importance sampling the optimal baseline differs from the reward mean (Greensmith et al., 2004; Weaver & Tao, 2013), and recent work derives baselines under simplifying assumptions (Hao et al., 2025) or tracks them globally (Xu & Ding, 2025), assuming importance weights independent of rewards. BICC conditions the ratio on

the opposite partition as privileged information (Pechyony & Vapnik, 2010; Lopez-Paz et al., 2015), and RCC corrects the baseline for the correlation that training induces (Appendix A).

3 PRELIMINARIES

3.1 REINFORCEMENT LEARNING FOR LLM REASONING

Consider a query q drawn from a task distribution $\mathcal{D}$ and an output sequence $o = (o_1, \ldots, o_T)$ with T tokens. The policy π_θ, parameterized by LLM weights θ, defines a distribution over outputs given q that factorizes autoregressively as $\pi_\theta(o|q) = \prod_{t=1}^{T} \pi_\theta(o_t|q, o_{<t})$, where $o_{<t} = (o_1, \ldots, o_{t-1})$ denotes the prefix. For reasoning tasks with verifiable answers, the reward function $R(o, q) \in \{0, 1\}$ is binary, where 1 is referred to as correct and 0 otherwise. The objective for training a reasoning model is to maximize expected reward over the policy weight θ:

$$J(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, o\sim\pi_\theta(\cdot|q)}[R(o, q)] \tag{1}$$

PPO. Proximal Policy Optimization (Schulman et al., 2017) updates the policy using importance sampling from a previous iterate $\pi_{\theta_{\text{old}}}$. At state $s_t = (q, o_{<t})$ with action $a_t = o_t$, the policy importance ratio is $\rho_t = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$. The clipped surrogate objective as an approximation to (1) is defined as

$$L^{\text{CLIP}}(\theta) = \mathbb{E}\big[\min\big(\rho_t A_t, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)A_t\big)\big], \tag{2}$$

where ϵ is the clipping threshold. The advantage $A_t = Q(s_t, a_t) - V(s_t)$ measures how much better action a_t is than the average action, with the action-value function Q and the state-value function V estimated by a learned critic.

GRPO. Group Relative Policy Optimization (Guo et al., 2025) eliminates the critic model by estimating advantages from grouped samples. For each query q, GRPO samples G outputs $\{o_1, \ldots, o_G\}$ from $\pi_{\theta_{\text{old}}}$, where $o_i = (o_{i,1}, \ldots, o_{i,T_i})$ has T_i tokens and receives reward $r_i = R(o_i, q)$. The token-level objective of GRPO averages the clipped surrogate over the tokens of each output and over the group:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i} L_{i,t}^{\text{CLIP}}\right], \tag{3}$$

where $L_{i,t}^{\text{CLIP}} = \min\big(\rho_{i,t}A_i, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_i\big)$ and $\rho_{i,t} = \pi_\theta(o_{i,t}|q, o_{i,<t})/\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})$. The advantage A_i is constant across all tokens within o_i and computed via group-wise standardization:

$$A_i = \frac{r_i - \mu}{\sigma}, \quad \mu = \frac{1}{G}\sum_{j=1}^{G} r_j, \quad \sigma = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu)^2}. \tag{4}$$

The group mean μ replaces the critic as a baseline, and division by σ normalizes the advantage scale across queries of varying difficulty.

3.2 POLICY GRADIENT WITH VARIANCE REDUCTION

The policy gradient theorem (Williams, 1992; Sutton et al., 1999) expresses the gradient of the expected reward as

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[R(o, q)\,\nabla_\theta \log \pi_\theta(o|q)] = \mathbb{E}\left[\tfrac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)} R(o, q)\,\nabla_\theta \log \pi_\theta(o|q)\right]. \tag{5}$$

Subtracting a baseline b reduces variance and leaves the expected gradient unchanged, and under importance sampling with weight $w = \pi_\theta/\pi_{\theta_{\text{old}}}$, minimizing the variance over b yields the optimal baseline (derivation in Appendix B.1):

$$\nabla_\theta J = \mathbb{E}\left[\tfrac{\pi_\theta}{\pi_{\theta_{\text{old}}}}(R - b)\,\nabla_\theta \log \pi_\theta\right], \qquad b^* = \frac{\mathbb{E}[Rw^2]}{\mathbb{E}[w^2]}. \tag{6}$$

When $w \equiv 1$ the baseline reduces to $b^* = \mathbb{E}[R]$, the group mean of GRPO, which is optimal only when importance weights are independent of rewards (Kakade & Langford, 2002; Schulman et al., 2015a).

4 ANOTHER LOOK AT GRPO AND VARIANCE REDUCTION

4.1 CONTRASTIVE REFORMULATION OF GRPO OBJECTIVE

Within each group, correct and incorrect outputs form two partitions with distinct optimization roles (Mroueh et al., 2025; Mroueh, 2025; Li et al., 2025), and we derive an equivalent form of the GRPO objective under binary rewards that makes the partition explicit.

Sample Partition. For each query q, we partition the group of outputs into two disjoint subsets based on their corresponding binary rewards $\{r_i\}_{i=1}^{G}$: $\mathcal{O}^+ = \{o_i \mid r_i = 1\}$ and $\mathcal{O}^- = \{o_i \mid r_i = 0\}$, where $G^+ = |\mathcal{O}^+|$ and $G^- = |\mathcal{O}^-|$ are the cardinalities and $\hat{p} = G^+/G$ denotes the sample percentage of positive rewards. Thus $\mu = \hat{p}$ and $\sigma^2 = \hat{p}(1-\hat{p})$. Substituting into Eq. (4) gives the per-group advantages

$$A^+ = \sqrt{\tfrac{1-\hat{p}}{\hat{p}}} > 0, \qquad A^- = -\sqrt{\tfrac{\hat{p}}{1-\hat{p}}} < 0. \tag{7}$$

Clipping acts asymmetrically in the sign of the advantage: for $A > 0$ the min yields $A \cdot \min(\rho, 1+\epsilon)$, and for $A < 0$ it yields $A \cdot \max(\rho, 1-\epsilon)$, so we define

$$\mathcal{C}_{\text{up}}(\rho) = \min(\rho, 1+\epsilon), \qquad \mathcal{C}_{\text{low}}(\rho) = \max(\rho, 1-\epsilon). \tag{8}$$

Contrastive Form. We treat ρ_i as a sequence-level ratio in this section, a surrogate for the token-level objective that is exact when clipping is inactive and a tight upper bound otherwise. Applying $\mathcal{C}_{\text{up}}$ to positive samples and $\mathcal{C}_{\text{low}}$ to negative samples, and using $\hat{p} \cdot A^+ = (1-\hat{p}) \cdot |A^-| = \sqrt{\hat{p}(1-\hat{p})} = \sigma_q$, the GRPO objective can be written as

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_q\left[\tfrac{G^+}{G} A^+ \bar{\rho}^+_{\text{clip}} + \tfrac{G^-}{G} A^- \bar{\rho}^-_{\text{clip}}\right] = \mathbb{E}_q\left[\sigma_q \cdot \left(\bar{\rho}^+_{\text{clip}} - \bar{\rho}^-_{\text{clip}}\right)\right], \tag{9}$$

where $\bar{\rho}^+_{\text{clip}} = \frac{1}{G^+}\sum_{i\in\mathcal{O}^+} \mathcal{C}_{\text{up}}(\rho_i^+)$ and $\bar{\rho}^-_{\text{clip}} = \frac{1}{G^-}\sum_{j\in\mathcal{O}^-} \mathcal{C}_{\text{low}}(\rho_j^-)$ are the average clipped ratios. Expanding the difference into pairwise form gives

$$\bar{\rho}^+_{\text{clip}} - \bar{\rho}^-_{\text{clip}} = \frac{1}{G^+G^-}\sum_{i=1}^{G^+}\sum_{j=1}^{G^-}\left(\mathcal{C}_{\text{up}}(\rho_i^+) - \mathcal{C}_{\text{low}}(\rho_j^-)\right), \tag{10}$$

and substituting back yields

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_q\left[\sum_{i=1}^{G^+}\sum_{j=1}^{G^-} A_{ij}\left(\mathcal{C}_{\text{up}}(\rho_i^+) - \mathcal{C}_{\text{low}}(\rho_j^-)\right)\right], \tag{11}$$

where $A_{ij} = \sigma_q/(G^+G^-)$ is shared by all pairs. GRPO therefore performs contrastive optimization over all positive-negative pairs, yet each ratio $\rho_i = \pi_\theta(o_i|q)/\pi_{\theta_{\text{old}}}(o_i|q)$ conditions the policy on the query alone, so the other samples of the group stay unseen. Wu et al. (2025) present a related contrastive view that removes the importance ratio to draw a DPO-style connection, whereas our reformulation preserves the clipping structure through $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{low}}$.

4.2 APPROXIMATING OPTIMAL BASELINE UNDER IMPORTANCE SAMPLING

Log-Probability Formulation. Direct computation of the importance weight $w = \pi_\theta/\pi_{\theta_{\text{old}}}$ in Eq. (6) is numerically unstable for long sequences, so we work with the log-probability difference

$$\delta(x, y) = \log \pi_\theta(y|x) - \log \pi_{\theta_{\text{old}}}(y|x), \tag{12}$$

so that $w = e^\delta$. Under the trust region between π_θ and $\pi_{\theta_{\text{old}}}$, $|\delta|$ is small and the first-order expansion gives $w \approx 1 + \delta$ and $w^2 \approx 1 + 2\delta$.

Corrected Baseline. Substituting the expansion into Eq. (6) gives

$$b^* = \frac{\mathbb{E}[Rw^2]}{\mathbb{E}[w^2]} \approx \frac{\mathbb{E}[R] + 2\mathbb{E}[R\delta]}{1 + 2\mathbb{E}[\delta]}. \tag{13}$$

Expanding to first order in $\mathbb{E}[\delta]$ and $\mathbb{E}[R\delta]$, the $\mathbb{E}[R]\,\mathbb{E}[\delta]$ cross-terms cancel, leaving

$$b^* \approx \mathbb{E}[R] + 2\big(\mathbb{E}[R\delta] - \mathbb{E}[R]\,\mathbb{E}[\delta]\big) = \mathbb{E}[R] + 2 \cdot \mathrm{Cov}(R, \delta). \tag{14}$$

When $\mathrm{Cov}(R, \delta) > 0$, the correction increases the baseline, preventing high-confidence correct samples from dominating the gradient.

Table 1: The table applies BICC to GRPO variants, where the shaded row of each method is the bilateral form and the change appears in green.

Ratio ρ_i	Objective $\mathcal{L}_i$
GRPO (Guo et al., 2025)	
$\frac{\pi_\theta(o_i\|q)}{\pi_{\theta_{\text{old}}}(o_i\|q)}$	$\min\big(\rho_i A_i, \text{clip}(\rho_i, 1\pm\epsilon)A_i\big)$
$\frac{\pi_\theta(o_i\|q, \mathcal{O}_i^{\mp})}{\pi_{\theta_{\text{old}}}(o_i\|q)}$	$\min\big(\rho_i A_i, \text{clip}(\rho_i, 1\pm\epsilon)A_i\big)$
Dr.GRPO (Liu et al., 2025)	
$\frac{\pi_\theta(o_i\|q)}{\pi_{\text{ref}}(o_i\|q)}$	$\min\big(\rho_i A_i, \text{clip}(\rho_i, 1\pm\epsilon)A_i\big) - \beta\mathbb{D}_{\text{KL}}$
$\frac{\pi_\theta(o_i\|q, \mathcal{O}_i^{\mp})}{\pi_{\text{ref}}(o_i\|q)}$	$\min\big(\rho_i A_i, \text{clip}(\rho_i, 1\pm\epsilon)A_i\big) - \beta\mathbb{D}_{\text{KL}}$
DAPO (Yu et al., 2025)	
$\frac{\pi_\theta(o_i\|q)}{\pi_{\theta_{\text{old}}}(o_i\|q)}$	$\min\big(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon_l, 1+\epsilon_h)A_i\big)$
$\frac{\pi_\theta(o_i\|q, \mathcal{O}_i^{\mp})}{\pi_{\theta_{\text{old}}}(o_i\|q)}$	$\min\big(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon_l, 1+\epsilon_h)A_i\big)$
GSPO (Zheng et al., 2025)	
$\exp\Big(\frac{1}{\lvert o_i\rvert}\sum_t \log\frac{\pi_\theta(o_{i,t}\|q)}{\pi_{\theta_{\text{old}}}(o_{i,t}\|q)}\Big)$	$\rho_i^{1/\lvert o_i\rvert}\cdot \text{sg}(\rho_i)^{1-1/\lvert o_i\rvert}\cdot A_i$
$\exp\Big(\frac{1}{\lvert o_i\rvert}\sum_t \log\frac{\pi_\theta(o_{i,t}\|q, \mathcal{O}_i^{\mp})}{\pi_{\theta_{\text{old}}}(o_{i,t}\|q)}\Big)$	$\rho_i^{1/\lvert o_i\rvert}\cdot \text{sg}(\rho_i)^{1-1/\lvert o_i\rvert}\cdot A_i$

$\mathcal{O}_i^{\mp}$: opposite-partition samples for o_i. $\text{sg}(\cdot)$: stop-gradient. $\mathbb{D}_{\text{KL}}$: KL divergence from reference policy.

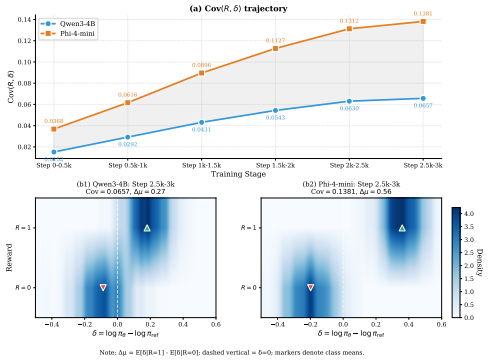


Figure 2: $\text{Cov}(R, \delta)$ rises throughout training on both models, and δ at steps 2.5k to 3k separates by reward with $\Delta_\mu = 0.27$ on Qwen3-4B and 0.56 on Phi-4-mini. The axis label writes the rollout policy $\pi_{\theta_{\text{old}}}$ as π_{ref}.

5 METHOD

As Figure 1 illustrates, BICC conditions each sample on opposite-partition outputs (§5.1) and RCC adjusts the baseline by the reward-confidence covariance (§5.2).

5.1 BILATERAL CONTEXT CONDITIONING

In the pairwise form of Eq. (11) neither member of a pair (o_i^+, o_j^-) observes the other during learning. Bilateral Context Conditioning (BICC) opens that information flow while preserving the contrastive structure, by concatenating the query with opposite-partition samples:

$$x^+ = [q; \mathcal{O}^-], \qquad x^- = [q; \mathcal{O}^+].$$

When evaluating correct solutions, the policy observes failed reasoning traces via x^+, and when evaluating incorrect ones, it observes successful traces via x^-.

Budgeted Context. The context contains full sampled outputs, reasoning traces included, since the contrast resides in how each attempt proceeds. A fixed fraction of the maximum context length, 0.4 of 2048 tokens in our setting, is reserved for it, so the augmented sequence stays within the length the baseline processes. For each sample, opposite-partition outputs are drawn uniformly without replacement and appended until the next would exceed the budget, and the remaining candidates are dropped instead of being truncated mid-solution. The draw is repeated at each epoch, and with rollouts of 385 tokens on average for Qwen3-4B and 420 for Phi-4-mini, about two opposite-partition samples fit per conditioned sample. When $G^+ \neq G^-$, the smaller partition supplies its samples to several members of the larger one, so each sample receives a context whenever both partitions are non-empty. From here on, the symbol $\mathcal{O}^{\mp}$ denotes the budgeted context of each sample.

Conditioned Importance Sampling Ratios. BICC replaces the importance ratio with its conditioned form, which lets information flow directly across partitions:

$$\begin{aligned} \rho_i^+ &= \frac{\pi_\theta(o_i \mid x^+)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, \quad o_i \in \mathcal{O}^+, \\ \rho_j^- &= \frac{\pi_\theta(o_j \mid x^-)}{\pi_{\theta_{\text{old}}}(o_j \mid q)}, \quad o_j \in \mathcal{O}^-. \end{aligned} \tag{15}$$

Algorithm 1 BICC with RCC

Require: Dataset $\mathcal{D}$, policy π_θ, old policy $\pi_{\theta_{\text{old}}}$, group size G

for each query $q \in \mathcal{D}$ **do**
  # Sampling and partitioning
  Sample $\{o_i\} \sim \pi_{\theta_{\text{old}}}(\cdot|q)$ with rewards $\{r_i\}$; $\mathcal{O}^+ = \{o_i : r_i{=}1\}$, $\mathcal{O}^- = \{o_i : r_i{=}0\}$
  # BICC: Bilateral Context Conditioning
  for $o_i \in \mathcal{O}^+ \cup \mathcal{O}^-$ **do**
    Draw $\mathcal{O}_i^{\mp}$ from the opposite partition within the token budget; $x^c = [q; \mathcal{O}_i^{\mp}]$
    $\rho_i^c = \pi_\theta(o_i|x^c)/\pi_{\theta_{\text{old}}}(o_i|q)$; $\delta_i = \log \pi_\theta(o_i|q) - \log \pi_{\theta_{\text{old}}}(o_i|q)$
  end for
  # RCC: Reward-Confidence Correction
  $\bar{R} = \frac{1}{G}\sum_i r_i$, $\bar{\delta} = \frac{1}{G}\sum_i \delta_i$, $\widehat{\text{Cov}} = \frac{1}{G}\sum_i (r_i - \bar{R})(\delta_i - \bar{\delta})$; $A_i^{\text{RCC}} = r_i - \bar{R} - 2\widehat{\text{Cov}}$
  # Policy update
  Update θ using $\{\rho_i^c, A_i^{\text{RCC}}\}$ with clipped surrogate objective
end for
return π_θ

Opposite-partition samples serve as privileged information (Lopez-Paz et al., 2015) available only during training, enabling contrastive learning with zero inference overhead.

Conditioned GRPO Objective. Replacing the independent ratios in Eq. (9) with conditioned ones yields:

$$J_{\text{cGRPO}}(\theta) = \mathbb{E}_q\left[\sigma_q \cdot \left(\bar{\rho}^+_{\text{clip}} - \bar{\rho}^-_{\text{clip}}\right)\right], \tag{16}$$

where $\bar{\rho}^+_{\text{clip}} = \frac{1}{G^+}\sum_{i\in\mathcal{O}^+} \mathcal{C}_{\text{up}}(\rho_i^+)$ and $\bar{\rho}^-_{\text{clip}} = \frac{1}{G^-}\sum_{j\in\mathcal{O}^-} \mathcal{C}_{\text{low}}(\rho_j^-)$. The objective structure remains identical to standard GRPO; only the conditioning of the trainable policy changes.

Ratio Decomposition and Trust Region. The conditioned ratio decomposes as $\rho_i^c = w_i \cdot \rho_i$, where

$$w_i = \frac{\pi_\theta(o_i \mid q, \mathcal{O}^{\mp})}{\pi_\theta(o_i \mid q)} \tag{17}$$

is a conditioning weight measuring the probability shift from observing opposite-partition samples, and cGRPO reduces to GRPO when $w_i = 1$ for all samples. The adaptation applies across the GRPO family by replacing $\pi_\theta(\cdot|q)$ with $\pi_\theta(\cdot|q, \mathcal{O}_i^{\mp})$ and keeping the remaining design choices of each variant, as Table 1 shows. Since numerator and denominator condition on different inputs, ρ_i^+ is a context-augmented surrogate and not a likelihood ratio on a common support, with the augmented numerator as privileged information (Lopez-Paz et al., 2015). The unaugmented denominator keeps the update referenced to the original sampling distribution, the clip thresholds of Eq. (8) bound ρ_i^+ independently of its interpretation and cap the per-step contribution of each sample. Appendix B.4 derives that w_i is bounded under standard Lipschitz assumptions.

5.2 Reward-Confidence Correction

The group mean is the optimal baseline only when importance weights are independent of rewards, and the model assigns higher probability to the outputs it considers correct, which induces a correlation between rewards and confidence. Figure 2 provides the evidence: $\text{Cov}(R, \delta)$ where $\delta = \log \pi_\theta - \log \pi_{\theta_{\text{old}}}$ increases monotonically throughout training, reaching 0.066 for Qwen3-4B and 0.138 for Phi-4-mini. The distribution of δ shows clear separation between correct and incorrect samples, with mean differences of 0.27 and 0.56 respectively.

Covariance-Based Correction. Section 4.2 derived the approximation of the variance-minimizing baseline under importance sampling:

$$b^* \approx \mathbb{E}[R] + 2 \cdot \text{Cov}(R, \delta) \tag{18}$$

The term $2 \cdot \text{Cov}(R, \delta)$ raises the baseline when the model assigns higher probability to correct outputs, which keeps high-confidence correct samples from dominating the gradient, and we call the adjustment Reward-Confidence Correction (RCC).

Corrected Advantage. We estimate the covariance within each sampled group using sample statistics:

$$\widehat{\text{Cov}}(R, \delta) = \frac{1}{G}\sum_{i=1}^{G}(r_i - \bar{R})(\delta_i - \bar{\delta}), \tag{19}$$

Table 2: The table reports Pass@1 (%) on mathematical reasoning benchmarks, averaged over 32 evaluation seeds, for Qwen3-4B and Phi-4-mini. Shaded rows apply BICC to the base method above them, with subscripts giving the absolute change, bold the best per column, and † the compute-matched GRPO with an enlarged rollout budget; the Baseline row is the released checkpoint before RL training.

	Qwen3-4B-Instruct-2507				Phi-4-mini-instruct-3.8B			
Method	Math500	AMC 23	AIME24	AIME25	Math500	AMC 23	AIME24	AIME25
Baseline	89.2	69.4	53.8	45.4	72.3	36.5	9.8	4.3
Standard GRPO Baseline								
GRPO ($G = 2$)	90.6	70.2	53.5	45.8	75.8	39.2	10.2	4.5
BICC GRPO	91.0↑0.4	71.4↑1.2	54.2↑0.7	46.2↑0.4	76.4↑0.6	40.5↑1.3	10.5↑0.3	4.8↑0.3
GRPO ($G = 8$)	91.4	72.8	54.0	46.4	76.2	42.5	10.4	4.5
GRPO ($G = 9$)†	91.5↑0.1	73.0↑0.2	54.1↑0.1	46.5↑0.1	76.4↑0.2	42.7↑0.2	10.5↑0.1	4.5 − 0.0
BICC GRPO	92.2↑0.8	74.2↑1.4	54.6↑0.6	47.1↑0.7	78.1↑1.9	44.3↑1.8	11.2↑0.8	5.1↑0.6
Generalization to GRPO Family ($G = 8$)								
Dr.GRPO (Liu et al., 2025)	91.5	72.9	53.9	46.5	76.4	42.6	10.5	4.6
BICC Dr.GRPO	92.1↑0.6	74.0↑1.1	54.5↑0.6	47.0↑0.5	78.2↑1.8	44.2↑1.6	11.1↑0.6	4.5↓0.1
ASPO (Wang et al., 2025)	91.6	72.6	54.1	46.3	76.5	42.4	10.5	4.5
BICC ASPO	91.9↑0.3	73.9↑1.3	54.7↑0.6	46.9↑0.6	78.3↑1.8	44.1↑1.7	11.3↑0.8	5.1↑0.6
GMPO (Zhao et al., 2025)	91.8	73.5	54.2	46.5	76.0	42.1	10.3	4.4
BICC GMPO	92.5↑0.7	74.6↑1.1	54.9↑0.7	47.2↑0.7	77.8↑1.8	42.0↓0.1	11.0↑0.7	5.0↑0.6
DAPO (Yu et al., 2025)	92.5	74.0	55.2	47.5	77.5	43.8	11.5	4.9
BICC DAPO	**93.1**↑0.6	75.2↑1.2	**55.8**↑0.6	48.0↑0.5	79.0↑1.5	**45.2**↑1.4	12.1↑0.6	**5.4**↑0.5
GSPO (Zheng et al., 2025)	92.3	74.2	55.0	47.7	77.8	43.5	11.6	4.8
BICC GSPO	92.9↑0.6	**75.4**↑1.2	55.5↑0.5	**48.3**↑0.6	**79.2**↑1.4	44.9↑1.4	**12.3**↑0.7	5.3↑0.5

where $\bar{R} = \frac{1}{G}\sum_i r_i$ and $\bar{\delta} = \frac{1}{G}\sum_i \delta_i$. The corrected advantage subtracts twice the estimate from the centered reward:

$$A_i^{\mathrm{RCC}} = r_i - \bar{R} - 2 \cdot \widehat{\mathrm{Cov}}(R, \delta) \tag{20}$$

The advantage omits the σ-normalization of GRPO: with binary rewards $\sigma = \sqrt{\hat{p}(1-\hat{p})} \leq 0.5$, so the division magnifies the advantage by at least a factor of 2 on top of the covariance correction, which already scales with the reward variance of the group. We therefore retain the unnormalized form in A_i^{RCC}, an effect we confirm empirically (Appendix B.5). The corrected advantage further satisfies $\sum_{i=1}^{G} A_i^{\mathrm{RCC}} = -2G \cdot \widehat{\mathrm{Cov}}(R, \delta) < 0$ whenever $\widehat{\mathrm{Cov}}(R, \delta) > 0$, shifting the policy update toward a globally cautious regime that shifts the gradient contributions downward while preserving their relative ranking, consistent with the variance reduction reported in Section 6.3. The computational overhead is small: $\{r_i\}$ come from the reward step, $\{\log \pi_\theta(o_i|q)\}$ takes one forward pass of the unaugmented sequence, and the covariance estimate requires only $O(G)$ additional operations per query.

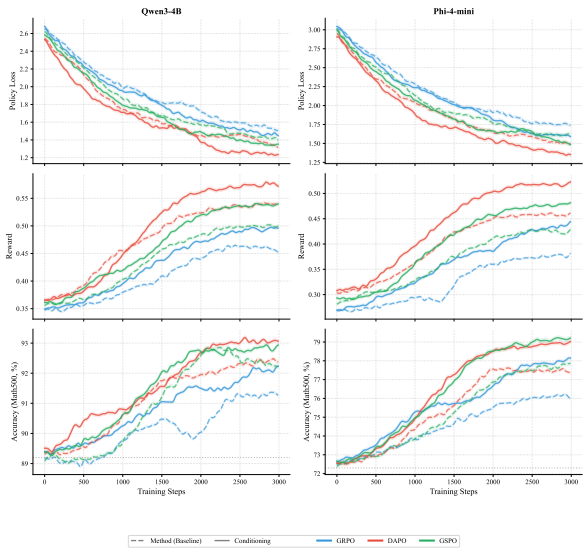
Algorithm 1 summarizes the procedure: BICC reshapes the importance ratio through the cross-partition context and RCC corrects the advantage through the covariance, so the two apply independently or jointly to GRPO and its variants, and at inference the model generates from q alone.

6 Experiments

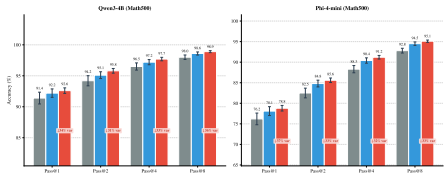
6.1 Experimental Setup

Models and Data. We evaluate on two instruction-tuned models of comparable scale, Qwen3-4B-Instruct-2507 and Phi-4-mini-instruct-3.8B, which reach strong baseline accuracy on mathematical reasoning while remaining tractable for extensive experimentation, and train on DAPO-Math-17k (Yu et al., 2025), approximately 17K problems with a ground-truth integer answer that supplies the binary reward.

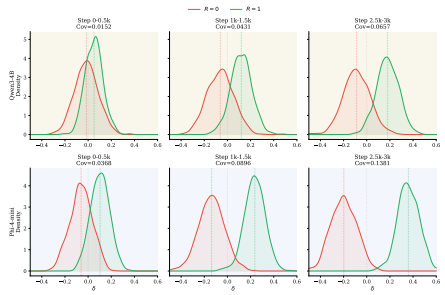
Baselines and Benchmarks. We apply BICC to GRPO and to five recent variants: Dr.GRPO uses a fixed reference policy in the denominator, DAPO employs asymmetric clipping bounds, and ASPO,



(a) Training dynamics; solid lines are conditioned variants, dashed lines their baselines.



(b) Pass@k on Math500, ± 1 SD over 32 evaluation seeds; red annotations give the variance reduction of BICC-GRPO-RCC over GRPO.



(c) Evolution of the δ distributions by reward.

Figure 3: The figure shows training behaviour on Qwen3-4B and Phi-4-mini, namely (a) convergence under BICC, (b) accuracy and gradient variance under RCC, and (c) the reward-confidence separation behind RCC.

GMPO, and GSPO modify advantage estimation or probability aggregation. These variants share the group structure that the conditioning targets. Evaluation uses four benchmarks of increasing difficulty: Math500 (Hendrycks et al., 2021), AMC 2023 from the American Mathematics Competition (Cobbe et al., 2021), and AIME 2024/2025 from the American Invitational Mathematics Examination (Zhang & Math-AI, 2025).

Implementation. Training uses 4×A100 GPUs with batch size 8, global batch size 32, learning rate 1×10^{-6} in AdamW, group size $G = 8$, and $\epsilon = 0.2$ unless noted. The opposite-partition context follows Section 5.1, the Baseline row of Table 2 is the released checkpoint evaluated under our protocol before RL training, and results are Pass@1 averaged over 32 evaluation seeds.

6.2 Main Results

Table 2 reports the performance of BICC applied to GRPO and its variants. Among the 56 comparisons against the matched baseline, 54 favour BICC, the two exceptions decline by 0.1 points, and the mean gain is $+0.74$ points on Qwen3-4B and $+0.98$ on Phi-4-mini, with per-benchmark significance over 32 seeds in Table 9. Increasing the group size from $G = 2$ to $G = 8$ amplifies the effect, from +0.4 to +0.8 points on Math500 for Qwen3-4B and from +0.6 to +1.9 for Phi-4-mini, since a larger opposite partition represents more failure or success modes. Matching the training FLOPs of BICC by a larger rollout budget yields negligible gains for GRPO (Table 2, † row), so the improvement comes from the cross-partition information flow and not from compute. The larger gains on Phi-4-mini indicate that weaker base models benefit more from the explicit contrast. The gains persist at 14B, where they shrink to 0.5 to 0.9 points, and extend to Llama-3.1-8B-Instruct with 2.1 points on Math500 (Appendix D.4), while training on mathematics alone transfers zero-shot to logic and code with gains of 0.8 to 1.7 points (Appendix D.5).

The gains persist across Dr.GRPO, ASPO, DAPO, GMPO, and GSPO, so BICC addresses the group structure itself and not an artifact of one algorithm, and BICC-DAPO and BICC-GSPO reach the highest accuracy, 93.1% and 79.2% on Math500 for Qwen3-4B and Phi-4-mini. Figure 3a shows that conditioned variants converge at a comparable or faster rate than their baselines, and the longer context leaves training stable. Variance reduction comes from RCC: in Figure 3b, adding RCC on top of BICC-GRPO raises Pass@k at $k = 1$ to 8 and brings gradient variance 31 to 36% below GRPO on Qwen3-4B and 32 to 37% on Phi-4-mini.

Table 3: The table reports ablations on Math500 with $G = 8$, Pass@1 (%) averaged over 32 evaluation seeds, for (a) context allocation, (b) no conditioning, same-class conditioning at fixed context length, and the cross-class conditioning of BICC-GRPO, and (c) σ-normalization on top of RCC on Qwen3-4B, with accuracy and gradient variance reduction relative to GRPO.

(a) Context allocation ratio.

Ratio	Qwen3	Phi-4
20%	91.8	77.4
40%	**92.2**	**78.1**
60%	92.0	77.8

(b) Conditioning choice.

Type	Qwen3	Phi-4
None	91.4	76.2
Same-class	91.5	76.4
Cross-class	**92.2**	**78.1**

(c) σ-norm in RCC (Qwen3).

Config	Acc.	Var. Red.
BICC-GRPO	92.2	–
+RCC	**92.6**	34%
+RCC+σ	91.9	51%

6.3 Ablation and Analysis

Context Length Allocation. Table 3a varies the proportion of the maximum context length allocated to opposite-partition samples in BICC-GRPO with $G = 8$. Too little context limits the contrastive information and too much dilutes the query, so 40% balances contrastive information against dilution of the query. Performance stays stable across the range, and Appendix D.7 repeats the three ablations on AIME 24 with the same orderings. Phi-4-mini is the more sensitive model, with a gap of 0.7 points between the best and worst settings against 0.4 for Qwen3-4B, which suggests that models with weaker initial capability rely more heavily on the contrastive signal.

Cross-Class versus Same-Class Conditioning. To test whether the gains stem from cross-partition contrast or from extended context alone, we replace opposite-partition samples with same-partition ones at the same context length, which yields +0.1 points on Qwen3-4B and +0.2 on Phi-4-mini over GRPO (Table 3b), against +0.8 and +1.9 for cross-class conditioning, so the contrast between partitions drives the gain and the extra context does not.

Reward-Confidence Correlation. Section 5.2 motivates RCC by a positive $\mathrm{Cov}(R, \delta)$. Figure 3c shows the evolution of $\delta = \log \pi_\theta - \log \pi_{\theta_{\text{old}}}$ by reward, and the separation between the $R = 0$ and $R = 1$ distributions widens progressively, from a large overlap early in training to distinct clusters by steps 2.5k to 3k. Phi-4-mini exhibits larger separation ($\Delta_\mu = 0.56$) than Qwen3-4B ($\Delta_\mu = 0.27$), indicating that the weaker model develops a more pronounced confidence gap. Table 3c ablates σ-normalization on top of RCC: removing the division yields the best accuracy (92.6%) with 34% variance reduction, and retaining it over-regularizes to 91.9% despite further reduction. With RCC, BICC-GRPO reaches 34% lower gradient variance than GRPO on Qwen3-4B while its accuracy rises from 92.2 to 92.6 (Table 3c).

7 Conclusion

We presented a contrastive reformulation of GRPO in which the correct and incorrect outputs of a group inform each other during training. On this reformulation we built Bilateral Context Conditioning, which opens the cross-partition information flow as privileged information, and Reward-Confidence Correction, which corrects the baseline by the reward-confidence covariance. Both integrate into existing GRPO variants with minimal overhead and improve 54 of 56 comparisons on mathematical reasoning benchmarks by 0.3 to 1.9 points, with larger gains on weaker base models. The gains transfer zero-shot to logic and code (Appendix D.5) and persist across a third model family and a second corpus (Appendices D.4 and D.6).

Limitations. The reformulation partitions a group by a binary outcome, so the method applies to tasks with an outcome checker, such as mathematics, code with unit tests, and constraint satisfaction. A continuous reward requires a soft partition, by thresholding at the group median or by weighting each pairwise term of Eq. (11) with the reward difference, which we leave to future work. Groups with a single outcome fall back to GRPO, at 12% of queries on Qwen3-4B and 18% on Phi-4-mini, and training cost rises by 5% in step time and 12% in peak memory (Appendix C.2.3). The deployed policy generates from q alone, so the reported numbers are measured on the bare query, and Appendix E.5 discusses reward quality and long solutions. Tool use lies outside the evaluated scope, and transfer to it remains a hypothesis for future verification.

AI USE STATEMENT

In this work, we used generative AI tools only to polish the writing of the manuscript. We have not used generative AI tools for developing the theoretical model or conceptual framework, formulating mathematical claims, designing the methodology or experiments, implementing the method, or interpreting the results, and the remaining tasks with required disclosure are not applicable to this work. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies mathematical reasoning on publicly available datasets and benchmarks. It involves neither human subjects nor personal or sensitive data, and we consider the work compliant with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The base models, the training data, the baselines, the benchmarks, and the evaluation protocol are specified in §6.1, and Appendix C lists the hyperparameters, the dataset construction, and the computational overhead in Appendices C.1, C.2, and C.2.3. The contrastive reformulation of GRPO, the variance-minimizing baseline and its first-order correction, and the boundedness of the conditioned importance ratio are derived in Appendices B.2, B.3, and B.4, with the scale analysis of the normalization in Appendix B.5. The supplementary experiments appear in Appendix D, and the source code is available in the repository linked in the abstract.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov):1471–1530, 2004.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yaru Hao, Li Dong, Xun Wu, Shaohan Huang, Zewen Chi, and Furu Wei. On-policy rl with optimal reward baseline. *arXiv preprint arXiv:2505.23585*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Yu Li, Tian Lan, and Zhengling Qi. Inspo: Unlocking intrinsic self-reflection for llm preference optimization. *arXiv preprint arXiv:2512.23126*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, Jiri Navratil, Jerret Ross, and Jesus Rios. Revisiting group relative policy optimization: Insights into on-policy and off-policy training. *arXiv preprint arXiv:2505.22257*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Dmitry Pechyony and Vladimir Naumovich Vapnik. On the theory of learning with privileged information. In *Neural Information Processing Systems*, 2010. URL https://api.semanticscholar.org/CorpusID:13217651.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Amir Saeidi, Shivanshu Verma, Md Nayem Uddin, and Chitta Baral. Insights into alignment: Evaluating dpo and its variants across multiple tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 409–421, 2025.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Yuanzhu Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. *arXiv preprint arXiv:1301.2315*, 2013.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, et al. It takes two: Your grpo is secretly dpo. *arXiv preprint arXiv:2510.00977*, 2025.

Zhongwen Xu and Zihan Ding. Single-stream policy optimization. *arXiv preprint arXiv:2509.13232*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Related Work

Reinforcement Learning for LLM Alignment. The application of reinforcement learning to large language models has evolved from early RLHF approaches (Christiano et al., 2017; Ouyang et al., 2022) to recent reasoning-focused training. Chain-of-thought prompting (Wei et al., 2022) demonstrated that intermediate reasoning steps substantially improve performance, motivating RL methods such as those used in OpenAI's o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025). Contrastive approaches such as DPO (Rafailov et al., 2023) and IPO (Saeidi et al., 2025) frame alignment as preference optimization, bypassing reward modeling (Ethayarajh et al., 2024; Azar et al., 2024). GRPO (Shao et al., 2024) addresses the computational overhead of PPO by eliminating the critic network and computing advantages through group-wise normalization. Variants include Dr.GRPO (Liu et al., 2025) (fixed reference policy), DAPO (Yu et al., 2025) (asymmetric clipping), and GSPO (Zheng et al., 2025) (geometric averaging). Existing methods evaluate each sample independently and leave the contrast between correct and incorrect outputs unused. BICC addresses this gap by conditioning on the opposite partition and integrates into GRPO and its variants.

Variance Reduction in Policy Gradients. Variance reduction through baselines is classical in policy gradient methods (Williams, 1992). The optimal baseline depends on the sampling distribution and generally differs from the simple reward mean (Greensmith et al., 2004; Weaver & Tao, 2013). Generalized Advantage Estimation (Schulman et al., 2015b) provides a practical bias-variance tradeoff, while RLOO (Ahmadian et al., 2024) revisits REINFORCE-style optimization. Recent work derives theoretically optimal baselines under simplifying assumptions (Hao et al., 2025) or replaces group-based estimation with persistent global trackers (Xu & Ding, 2025). Existing approaches assume independence between importance weights and rewards, which is increasingly violated as training progresses. Our method accounts for the resulting correlation via a covariance correction term.

B Theoretical Derivations

B.1 Optimal Baseline Derivation

We consider the objective of maximizing the expected reward for reasoning tasks, defined as $J(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, o\sim\pi_\theta(\cdot|q)}[R(o,q)]$, where $R(o,q) \in \{0,1\}$. To reduce the variance of the policy gradient estimator for a specific group i, we introduce a baseline b_i. Since the expected gradient is invariant to b_i, minimizing the gradient estimator's variance is equivalent to minimizing the expected squared weighted advantage with respect to b_i.

Proof. Let w_j denote the importance weight and R_j the reward for the j-th sample. We define the loss function $L(b_i)$ as the expected squared weighted advantage:

$$L(b_i) = \mathbb{E}\left[\left((R_j - b_i)w_j\right)^2\right] \tag{21}$$

Expanding the quadratic term and applying the linearity of expectation, we obtain:

$$\begin{aligned} L(b_i) &= \mathbb{E}\left[R_j^2 w_j^2 - 2R_j w_j^2 b_i + b_i^2 w_j^2\right] \\ &= \mathbb{E}[R_j^2 w_j^2] - 2b_i\mathbb{E}[R_j w_j^2] + b_i^2\mathbb{E}[w_j^2] \end{aligned} \tag{22}$$

To find the optimal baseline b_i^*, we take the derivative of $L(b_i)$ with respect to b_i and set it to zero:

$$\frac{\partial L}{\partial b_i} = -2\mathbb{E}[R_j w_j^2] + 2b_i\mathbb{E}[w_j^2] = 0 \tag{23}$$

Rearranging the terms yields:

$$b_i\mathbb{E}[w_j^2] = \mathbb{E}[R_j w_j^2] \tag{24}$$

Solving for b_i, we derive the optimal baseline:

$$b_i^* = \frac{\mathbb{E}[R_j w_j^2]}{\mathbb{E}[w_j^2]} \tag{25}$$

□

B.2 Contrastive Reformulation of GRPO

B.2.1 Advantage Simplification under Binary Rewards

Proposition 1. *Under binary rewards $r_i \in \{0, 1\}$, the group-normalized advantages reduce to:*

$$A^+ = \sqrt{\frac{1-\hat{p}}{\hat{p}}}, \quad A^- = -\sqrt{\frac{\hat{p}}{1-\hat{p}}} \tag{26}$$

where $\hat{p} = G^+/G$ is the proportion of correct samples.

Proof. From Eq. (4), the advantage is $A_i = (r_i - \mu)/\sigma$ with $\mu = \frac{1}{G}\sum_j r_j$ and $\sigma^2 = \frac{1}{G}\sum_j (r_j - \mu)^2$.

Under binary rewards, $\mu = G^+/G = \hat{p}$. The variance becomes:

$$\sigma^2 = \frac{1}{G}\left[G^+(1-\hat{p})^2 + G^-\hat{p}^2\right] = \hat{p}(1-\hat{p})^2 + (1-\hat{p})\hat{p}^2 = \hat{p}(1-\hat{p}) \tag{27}$$

For $r_i = 1$: $A^+ = (1-\hat{p})/\sqrt{\hat{p}(1-\hat{p})} = \sqrt{(1-\hat{p})/\hat{p}}$.

For $r_i = 0$: $A^- = -\hat{p}/\sqrt{\hat{p}(1-\hat{p})} = -\sqrt{\hat{p}/(1-\hat{p})}$. □

Corollary 1. *The advantages satisfy $\hat{p} \cdot A^+ = (1-\hat{p}) \cdot |A^-| = \sqrt{\hat{p}(1-\hat{p})} = \sigma_q$.*

B.2.2 Asymmetric Clipping Behavior

Lemma 1. *The clipped objective $L_i^{CLIP} = \min\left(\rho_i A_i, clip(\rho_i, 1-\epsilon, 1+\epsilon)A_i\right)$ satisfies:*

$$L_i^{CLIP} = \begin{cases} A_i \cdot \mathcal{C}_{up}(\rho_i) & \text{if } A_i > 0 \\ A_i \cdot \mathcal{C}_{low}(\rho_i) & \text{if } A_i < 0 \end{cases} \tag{28}$$

where $\mathcal{C}_{up}(\rho) = \min(\rho, 1+\epsilon)$ and $\mathcal{C}_{low}(\rho) = \max(\rho, 1-\epsilon)$.

Proof. When $A_i > 0$, both terms in the min are positive. The clipping only activates when $\rho_i > 1+\epsilon$, capping the contribution at $(1+\epsilon)A_i$. Thus $L_i^{\text{CLIP}} = A_i \cdot \min(\rho_i, 1+\epsilon)$.

When $A_i < 0$, both terms are negative, so min selects the more negative value. The clipping activates when $\rho_i < 1-\epsilon$, yielding $L_i^{\text{CLIP}} = A_i \cdot \max(\rho_i, 1-\epsilon)$. □

B.2.3 Derivation of the Contrastive Form

Theorem 1. *Under binary rewards, the GRPO objective can be written as:*

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_q\left[\sigma_q \cdot \left(\bar{\rho}^+_{clip} - \bar{\rho}^-_{clip}\right)\right] \tag{29}$$

where $\bar{\rho}^+_{clip} = \frac{1}{G^+}\sum_{i\in\mathcal{O}^+} \mathcal{C}_{up}(\rho_i^+)$ and $\bar{\rho}^-_{clip} = \frac{1}{G^-}\sum_{j\in\mathcal{O}^-} \mathcal{C}_{low}(\rho_j^-)$.

Proof. Separating the GRPO objective by partition and applying Lemma 1:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_q\left[\frac{1}{G}\left(\sum_{i\in\mathcal{O}^+} A^+\mathcal{C}_{\text{up}}(\rho_i^+) + \sum_{j\in\mathcal{O}^-} A^-\mathcal{C}_{\text{low}}(\rho_j^-)\right)\right] \tag{30}$$

$$= \mathbb{E}_q\left[\hat{p}\cdot A^+ \cdot \bar{\rho}^+_{\text{clip}} + (1-\hat{p})\cdot A^- \cdot \bar{\rho}^-_{\text{clip}}\right] \tag{31}$$

By Corollary 1, $\hat{p}\cdot A^+ = \sigma_q$ and $(1-\hat{p})\cdot A^- = -\sigma_q$:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_q\left[\sigma_q \cdot \bar{\rho}^+_{\text{clip}} - \sigma_q \cdot \bar{\rho}^-_{\text{clip}}\right] = \mathbb{E}_q\left[\sigma_q \cdot \left(\bar{\rho}^+_{\text{clip}} - \bar{\rho}^-_{\text{clip}}\right)\right] \tag{32}$$

□

B.2.4 Pairwise Contrastive Form

Corollary 2. *The objective admits a pairwise formulation:*

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_q \left[\sum_{i=1}^{G^+} \sum_{j=1}^{G^-} A_{ij} \left(\mathcal{C}_{up}(\rho_i^+) - \mathcal{C}_{low}(\rho_j^-) \right) \right] \tag{33}$$

where $A_{ij} = \sigma_q / (G^+ G^-)$.

Proof. For any sets $\{a_i\}_{i=1}^m$ and $\{b_j\}_{j=1}^n$:

$$\frac{1}{m} \sum_{i=1}^{m} a_i - \frac{1}{n} \sum_{j=1}^{n} b_j = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (a_i - b_j) \tag{34}$$

Applying this identity to $\bar{\rho}^+_{\text{clip}} - \bar{\rho}^-_{\text{clip}}$ and multiplying by σ_q yields the result. □

The pairwise form shows GRPO implicitly optimizes over all (o_i^+, o_j^-) pairs, yet each ratio $\rho_i = \pi_\theta(o_i|q)/\pi_{\theta_{\text{old}}}(o_i|q)$ conditions only on query q and leaves the other samples unseen. We note that this contrastive reformulation treats ρ_i as a sequence-level ratio for analytical clarity; standard token-level GRPO clipping (Eq. (3)) is averaged over per-token ratios, which differs from clipping the sequence-level product but yields equivalent objectives in the unclipped regime and a tight surrogate otherwise.

B.3 Optimal Baseline under Importance Sampling

B.3.1 Variance-Minimizing Baseline

Theorem 2. *The baseline minimizing variance of the importance-weighted gradient estimator is:*

$$b^* = \frac{\mathbb{E}[Rw^2]}{\mathbb{E}[w^2]} \tag{35}$$

where $w = \pi_\theta / \pi_{\theta_{old}}$ *is the importance weight relative to the rollout-collecting policy.*

Proof. The gradient estimator is $\hat{g} = (R - b) \cdot w \cdot \psi$ where $\psi = \nabla_\theta \log \pi_\theta$. Since subtracting a baseline does not change the expected gradient, minimizing variance reduces to minimizing $f(b) = \mathbb{E}[(R - b)^2 w^2 \psi^2]$.

Setting $\frac{df}{db} = -2\mathbb{E}[(R - b)w^2\psi^2] = 0$ gives:

$$b^* = \frac{\mathbb{E}[Rw^2\psi^2]}{\mathbb{E}[w^2\psi^2]} \tag{36}$$

Under the standard approximation that ψ^2 is approximately independent of R and w (Greensmith et al., 2004), this simplifies to $b^* = \mathbb{E}[Rw^2]/\mathbb{E}[w^2]$, where the expectation is taken with respect to the sampling distribution $\pi_{\theta_{\text{old}}}(\cdot|q)$. □

When $w \equiv 1$ (on-policy), this reduces to $b^* = \mathbb{E}[R]$.

B.3.2 First-Order Approximation under Trust Region

Proposition 2. *Under the trust region constraint between* π_θ *and* $\pi_{\theta_{old}}$*, the optimal baseline admits:*

$$b^* \approx \mathbb{E}[R] + 2 \cdot \mathit{Cov}(R, \delta) \tag{37}$$

where $\delta(x, y) = \log \pi_\theta(y|x) - \log \pi_{\theta_{old}}(y|x)$.

Proof. With $w = e^{\delta}$, the trust region constraint enforces small $|\delta|$, so the first-order Taylor expansion gives $w \approx 1 + \delta$ and $w^2 \approx 1 + 2\delta$. Thus:

$$\mathbb{E}[Rw^2] \approx \mathbb{E}[R] + 2\mathbb{E}[R\delta], \tag{38}$$

$$\mathbb{E}[w^2] \approx 1 + 2\mathbb{E}[\delta]. \tag{39}$$

Expanding the ratio $b^* = \mathbb{E}[Rw^2]/\mathbb{E}[w^2]$ to first order in the small quantities $\mathbb{E}[\delta]$ and $\mathbb{E}[R\delta]$:

$$b^* \approx \big(\mathbb{E}[R] + 2\mathbb{E}[R\delta]\big)\big(1 - 2\mathbb{E}[\delta]\big) \tag{40}$$

$$= \mathbb{E}[R] + 2\mathbb{E}[R\delta] - 2\mathbb{E}[R]\mathbb{E}[\delta] + O(\delta^2) \tag{41}$$

$$= \mathbb{E}[R] + 2\big(\mathbb{E}[R\delta] - \mathbb{E}[R]\mathbb{E}[\delta]\big) + O(\delta^2) \tag{42}$$

$$= \mathbb{E}[R] + 2\,\mathrm{Cov}(R, \delta) + O(\delta^2). \tag{43}$$

The $\mathbb{E}[R]\mathbb{E}[\delta]$ cross-terms cancel exactly to first order, so the additional assumption $\mathbb{E}[\delta] \approx 0$ is unnecessary. □

B.3.3 Corrected Advantage

The corrected advantage for sample i is:

$$A_i^{\text{RCC}} = r_i - \bar{R} - 2 \cdot \widehat{\mathrm{Cov}}(R, \delta) \tag{44}$$

where $\bar{R} = \frac{1}{G}\sum_j r_j$ and $\widehat{\mathrm{Cov}}(R, \delta) = \frac{1}{G}\sum_j (r_j - \bar{R})(\delta_j - \bar{\delta})$.

When $\mathrm{Cov}(R, \delta) > 0$ (model assigns higher probability to correct outputs), the correction increases the baseline, preventing high-confidence correct samples from dominating the gradient. When the correlation is zero, RCC reduces to standard GRPO.

RCC omits σ-normalization because the covariance term provides adaptive scaling on its own, since groups with higher reward variance yield correspondingly larger $|\mathrm{Cov}(R, \delta)|$. Moreover, in the binary reward setting $\sigma = \sqrt{\hat{p}(1-\hat{p})} \leq 0.5$, so dividing by σ amplifies the advantage magnitude by at least a factor of 2; stacking this on top of the covariance correction destabilizes the gradient scale, an effect we confirm empirically in Appendix B.5.

B.3.4 Computational Overhead

RCC requires $O(G)$ additional operations per query: computing $\delta_i = \log \pi_\theta(o_i|q) - \log \pi_{\theta_{\text{old}}}(o_i|q)$ from the log-probability of the unaugmented sequence, which takes one forward pass included in the advantage phase of Table 10, then $\bar{R}$, $\bar{\delta}$, and the sample covariance, whose cost is negligible next to that pass.

B.4 Theoretical Analysis of Bilateral Conditioning

This appendix formalizes the per-step update bound sketched in Section 5.1. We establish (i) the precise sense in which ρ_i^+ functions as a context-augmented surrogate, (ii) how the unaugmented denominator $\pi_{\theta_{\text{old}}}(\cdot \mid q)$ keeps the per-step update referenced to the original sampling distribution, and (iii) a bound on the conditioning weight w_i under standard Lipschitz assumptions.

B.4.1 Status of the Conditioned Importance Ratio

Proposition 3. *The conditioned ratio $\rho_i^+ = \pi_\theta(o_i \mid x^+)/\pi_{\theta_{old}}(o_i \mid q)$ admits the multiplicative decomposition*

$$\rho_i^+ = w_i \cdot \rho_i, \qquad w_i = \frac{\pi_\theta(o_i \mid q, \mathcal{O}^-)}{\pi_\theta(o_i \mid q)}, \tag{45}$$

where $\rho_i = \pi_\theta(o_i \mid q)/\pi_{\theta_{old}}(o_i \mid q)$ is the standard importance ratio between π_θ and $\pi_{\theta_{old}}$ on the original query distribution, and w_i is a context-shift factor measuring the change in policy probability induced by conditioning on opposite-partition samples.

Proof. Since $x^+ = [q; \mathcal{O}^-]$ by construction, $\pi_\theta(o_i \mid q, \mathcal{O}^-)$ and $\pi_\theta(o_i \mid x^+)$ refer to the same conditional distribution. Multiplying and dividing by $\pi_\theta(o_i \mid q)$ then gives the decomposition. □

The decomposition isolates two distinct effects: ρ_i captures the parameter change from θ_{old} to θ on the original query distribution, while w_i captures the context shift induced by privileged information $\mathcal{O}^-$. Standard PPO/GRPO clipping guarantees apply to ρ_i; the role of w_i is analyzed below.

B.4.2 Per-Step Update Bound via the Unaugmented Denominator

Proposition 4. *The bilateral objective J_{cGRPO} in Eq. (16) produces a per-step update whose contribution from each sample is bounded above in magnitude by $|A_i| \cdot (1+\epsilon) \cdot w_i^{\max}$, where $w_i^{\max} = \sup_i w_i$ is the maximum conditioning weight in the group.*

Proof. The conditioned ratios ρ_i^+ and ρ_j^- share the denominator $\pi_{\theta_{\text{old}}}(o \mid q)$ of standard GRPO. Decomposing $\rho_i^+ = w_i \cdot \rho_i$ where $\rho_i = \pi_\theta(o_i \mid q)/\pi_{\theta_{\text{old}}}(o_i \mid q)$, the upper clip operator $\mathcal{C}_{\text{up}}(\rho_i^+) = \min(\rho_i^+, 1+\epsilon)$ enforces $\rho_i^+ \le 1+\epsilon$ on the saturating side. Combined with $\rho_i^+ = w_i\rho_i$, this gives $\rho_i \le (1+\epsilon)/w_i$, so each sample's contribution to the surrogate is bounded above by $|A_i| \cdot (1+\epsilon)$. □

This bound on the surrogate objective should not be confused with a KL trust region between $\pi_\theta(\cdot \mid q)$ and $\pi_{\theta_{\text{old}}}(\cdot \mid q)$: the conditioned ratio mixes augmented and unaugmented distributions, and clipping it does not directly constrain $D_{\text{KL}}(\pi_\theta(\cdot \mid q) \,\|\, \pi_{\theta_{\text{old}}}(\cdot \mid q))$. The bound we establish is on the per-step magnitude of the policy gradient contribution, which is the property that motivates clipping in PPO-family methods.

B.4.3 Boundedness of the Conditioning Weight

Theorem 3. *Suppose π_θ is L-Lipschitz in its conditioning input under a suitable embedding metric, the augmented context length $|\mathcal{O}^-|$ is bounded by M tokens, and the output length $|o_i|$ is bounded by T. Then there exists a constant $C(L, M, T) < \infty$ such that the conditioning weight satisfies*

$$\frac{1}{C} \le w_i \le C \qquad \textit{for all } i. \tag{46}$$

Proof sketch. Let $\Phi(\cdot)$ denote the policy's hidden-state representation of the conditioning input. Lipschitz continuity gives $\|\Phi(q, \mathcal{O}^-) - \Phi(q)\| \le L \cdot \|\Phi(\mathcal{O}^-)\| \le LM'$ for some M' depending on M. Since o_i is a sequence of $|o_i| \le T$ tokens, the sequence log-likelihood decomposes as $\log \pi_\theta(o_i \mid \cdot) = \sum_{t=1}^{|o_i|} \log \pi_\theta(o_{i,t} \mid \cdot, o_{i,<t})$. Each per-token log-probability is a smooth function of the hidden state with bounded gradient norm G_π, so

$$|\log w_i| = \left| \sum_{t=1}^{|o_i|} \log \frac{\pi_\theta(o_{i,t} \mid q, \mathcal{O}^-, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} \right| \le |o_i| \cdot G_\pi \cdot LM' \le T \cdot G_\pi \cdot LM' =: \log C.$$

Exponentiating yields the stated bound. □

The bound implies w_i does not introduce unbounded magnification or attenuation of ρ_i, so the composite ratio $\rho_i^+ = w_i\rho_i$ inherits the magnitude control properties of ρ_i.

B.4.4 Clipping as a Hard Bound on Update Magnitude

Proposition 5. *The clipping operators enforce one-sided bounds:*

$$\mathcal{C}_{up}(\rho_i^+) \le 1+\epsilon, \qquad \mathcal{C}_{low}(\rho_j^-) \ge 1-\epsilon, \tag{47}$$

saturating in the direction relevant to the corresponding sign of the advantage, regardless of whether ρ_i^+ or ρ_j^- admits a likelihood-ratio interpretation.

Proof. Direct from the definitions $\mathcal{C}_{\text{up}}(\rho) = \min(\rho, 1+\epsilon)$ and $\mathcal{C}_{\text{low}}(\rho) = \max(\rho, 1-\epsilon)$. The bounds are functions of the numerical value of ρ alone and do not require ρ to correspond to a Radon-Nikodym derivative between two distributions on a common support. Note that $\mathcal{C}_{\text{up}}$ provides only an upper bound (with $\mathcal{C}_{\text{up}}(\rho) \to 0$ as $\rho \to 0$), and $\mathcal{C}_{\text{low}}$ provides only a lower bound (with $\mathcal{C}_{\text{low}}(\rho) \to \infty$ as $\rho \to \infty$); each operator saturates in the direction that limits the gradient contribution of advantage-positive (resp. advantage-negative) samples. □

Combined with Theorem 3, this establishes that each sample's per-step contribution to the surrogate objective is uniformly bounded under bilateral conditioning, preserving the variance-bounding role of clipping that motivates PPO-family methods. This bound concerns the magnitude of the gradient contribution rather than the KL distance between π_θ and $\pi_{\theta_{\text{old}}}$ on the original query distribution; the latter is not directly enforced by clipping the conditioned ratio, as discussed in Section 5.1.

B.4.5 Empirical Validation of Boundedness

We empirically verify that w_i remains in a moderate range throughout training. Table 4 reports the mean and standard deviation of w_i across training stages on Qwen3-4B with $G = 8$.

Table 4: Evolution of conditioning weight w_i during training (Qwen3-4B, $G = 8$).

Partition	0–0.5k	0.5k–1k	1k–1.5k	1.5k–2k	2k–2.5k	2.5k–3k
Correct (w_i^+)	1.01 ± 0.03	1.04 ± 0.04	1.08 ± 0.06	1.10 ± 0.07	1.12 ± 0.08	1.14 ± 0.09
Incorrect (w_i^-)	0.99 ± 0.04	0.95 ± 0.05	0.91 ± 0.07	0.88 ± 0.09	0.86 ± 0.10	0.84 ± 0.11

The weights diverge from 1.0 in opposite directions for the two partitions, confirming the intended contrastive mechanism: observing failures amplifies confidence in correct solutions ($w_i^+ > 1$), while observing successes attenuates confidence in flawed ones ($w_i^- < 1$). Throughout training, w_i remains in $[0.6, 1.4]$ at the 99th percentile, well within the bounded regime predicted by Theorem 3.

B.5 Scale Analysis of σ-Normalization in RCC

This appendix justifies the omission of σ-normalization in the RCC advantage estimator, complementing the magnitude argument in Section 5.2.

B.5.1 Dimensional Analysis

Proposition 6. *The* RCC *correction term* $\widehat{\mathrm{Cov}}(R, \delta)$ *shares the same physical units as the centered reward* $r_i - \bar{R}$.

Proof. The reward $r_i \in \{0, 1\}$ has reward units, denoted $[\mathcal{R}]$. The log-probability shift $\delta_i = \log \pi_\theta(o_i \mid q) - \log \pi_{\theta_{\text{old}}}(o_i \mid q)$ is dimensionless ($[\not\mathcal{K}]$). The sample covariance

$$\widehat{\mathrm{Cov}}(R, \delta) = \tfrac{1}{G} \sum_i (r_i - \bar{R})(\delta_i - \bar{\delta})$$

has units $[\mathcal{R}] \cdot [\not\mathcal{K}] = [\mathcal{R}]$, identical to $r_i - \bar{R}$. The RCC advantage $A_i^{\text{RCC}} = (r_i - \bar{R}) - 2\widehat{\mathrm{Cov}}(R, \delta)$ is therefore dimensionally consistent without further normalization. □

In contrast, standard GRPO's σ-normalization $A_i = (r_i - \bar{R})/\sigma$ rescales reward-unit quantities by the standard deviation $\sigma = \sqrt{\hat{p}(1 - \hat{p})}$. In the binary reward setting, $\sigma \leq 0.5$, so σ-division *amplifies* the advantage magnitude by at least a factor of 2. Applied on top of the dimensionally consistent RCC correction, this yields

$$\frac{(r_i - \bar{R}) - 2\widehat{\mathrm{Cov}}(R, \delta)}{\sigma},$$

in which both the centered reward and the covariance correction are simultaneously magnified, producing gradient updates substantially larger than either correction alone would generate. The resulting over-amplification destabilizes the gradient scale and degrades empirical performance, as verified below.

B.5.2 Empirical Confirmation

Table 5 provides the complete ablation comparing four configurations on Qwen3-4B and Phi-4-mini.

The pattern agrees across both models: combining σ-normalization with RCC produces the largest variance reduction but the worst accuracy, confirming the over-amplification argument above. RCC

Table 5: Effect of σ-normalization combined with RCC on Math500 ($G = 8$).

Configuration	Qwen3-4B		Phi-4-mini	
	Acc. (%)	Var. Red. (%)	Acc. (%)	Var. Red. (%)
BICC-GRPO (no RCC)	92.2	–	78.1	–
+ RCC (no σ-norm)	**92.6**	34	**78.8**	37
+ RCC (with σ-norm)	91.9	51	77.6	53
GRPO + σ-norm only	91.4	0	76.2	0

with σ-normalization disabled achieves the best accuracy with substantial variance reduction, validating the dimensional analysis: the covariance correction shares units with the centered reward, so σ-division is redundant.

C Experimental Details

C.1 Hyperparameters

Tables 6 and 7 list the hyperparameter configurations.

Table 6: Hyperparameter configuration (shared across both models unless noted).

Hyperparameter	Value
Optimization	
Learning rate	1×10^{-6}
Optimizer (Adam β_1, β_2)	AdamW (0.9, 0.999)
Weight decay / Gradient clipping	0.01 / 1.0
LR schedule (warmup / total steps)	Cosine (100 / 3000)
Batch & Model	
Per-device batch size / GPUs / Global batch	8 / 4 / 32
Precision / Flash Attention	bfloat16 / v2
GRPO & BICC	
Group size G / Clipping ϵ	8 / 0.2
Context allocation ratio / Max tokens	0.4 / 2048
Generation (training / evaluation)	
Max new tokens	2048 / 4096
Temperature / Top-p	1.0, 1.0 / 0.6, 0.95

Table 7: Variant-specific hyperparameters.

Method	Parameter	Value
Dr.GRPO	KL coefficient β; Reference policy	0.01; Frozen initial
DAPO	Clip bounds (ϵ_l, ϵ_h)	(0.2, 0.28)
GMPO	Geometric mean exponent	$1/T_i$
GSPO	Aggregation; Stop-gradient	Geometric mean; Yes

C.2 Dataset and Evaluation

C.2.1 Training Data

The DAPO-Math-17k dataset (Yu et al., 2025) contains ∼17,000 mathematical problems with integer answers. The reward function assigns binary scores: $R(o, a^*) = \mathbf{1}[\text{extract}(o) = a^*]$, where extract$(\cdot)$ parses the final answer (matching "The answer is X" or "$\boxed{X}$") and a^* is the ground truth.

C.2.2 Evaluation Benchmarks and Protocol

Table 8: Evaluation benchmarks and sampling configuration.

Benchmark	Problems	Difficulty	Answer Type	Source
Math500	500	Competition	Integer/Expr	Hendrycks et al. (2021)
AMC 2023	40	Competition	Integer	AMC 10/12
AIME 2024	30	Olympiad	Integer (0–999)	AIME I/II
AIME 2025	30	Olympiad	Integer (0–999)	AIME I/II
Evaluation Configuration: Temperature 0.6, Top-p 0.95, Max tokens 4096, Seeds 0–31				

Pass@k Estimation. For Pass@k metrics (Figure 3b), we generate $n = 32$ samples per problem and use the unbiased estimator (Chen et al., 2021): Pass@$k = \mathbb{E}[1 - \binom{n-c}{k} / \binom{n}{k}]$ where c is the number of correct samples. Error bars give ± 1 standard deviation over the 32 evaluation seeds.

Statistical Significance. For pairwise comparisons between BICC variants and their baselines, we compute Wilcoxon signed-rank tests across the 32 evaluation seeds. Table 9 reports the standard deviation of Pass@1 accuracy and the resulting p-values for each benchmark.

Table 9: Per-benchmark standard deviation and significance tests for BICC-GRPO vs. GRPO ($G = 8$, 32 seeds).

	Qwen3-4B				Phi-4-mini			
Method	Math500	AMC 23	AIME 24	AIME 25	Math500	AMC 23	AIME 24	AIME 25
GRPO	91.4±0.35	72.8±0.52	54.0±0.81	46.4±0.93	76.2±0.41	42.5±0.58	10.4±0.85	4.5±0.96
BICC-GRPO	92.2±0.33	74.2±0.49	54.6±0.78	47.1±0.92	78.1±0.38	44.3±0.55	11.2±0.82	5.1±0.94
Δ (BICC)	+0.8	+1.4	+0.6	+0.7	+1.9	+1.8	+0.8	+0.6
p-value	$< 10^{-5}$	$< 10^{-4}$	< 0.05	< 0.05	$< 10^{-6}$	$< 10^{-5}$	< 0.05	< 0.05

The improvements on Math500 and AMC 23 reach significance at $p < 0.001$; AIME 24 and AIME 25 results, despite the inherently smaller benchmark size (30 problems each), remain significant at $p < 0.05$. The standard error of the mean over 32 seeds is approximately $\sigma/\sqrt{32}$, yielding standard errors of 0.06 to 0.07 on Math500 and 0.14 to 0.17 on AIME, well below the observed mean differences.

C.2.3 Computational Overhead

BICC adds computational cost in the advantage forward pass and, to a lesser extent, in the backward pass, while rollout generation and inference are unchanged. We provide phase-level measurements on 4×A100 GPUs with Qwen3-4B and $G = 8$.

Table 10: Per-step time and memory breakdown comparing GRPO and BICC-GRPO (Qwen3-4B, $G = 8$).

Phase	GRPO	BICC-GRPO	Overhead
Rollout generation (autoregressive)	70% of step time	70% of step time	+0%
Advantage forward pass	15% of step time	19% of step time	+27%
Backward pass + optimizer step	15% of step time	16% of step time	+7%
Total step time	1.00×	1.05×	+5%
Peak memory	1.00×	1.12×	+12%

Phase-Level Breakdown. The overall step-time overhead is modest because rollout generation, dominated by autoregressive decoding, is the wall-clock bottleneck and is unaffected by BICC.

The 12% peak memory increase comes from storing activations for the extended context during the advantage forward pass; activation checkpointing can reduce this at the cost of additional recomputation during the backward pass.

Absolute Cost. Table 11 reports the absolute figures behind the ratios of Table 10. The augmented forward pass runs teacher-forced and parallelizes across positions, so the extra tokens cost far less time than their count suggests, and the sequence length is capped by construction at the maximum the baseline already uses.

Table 11: Absolute training cost on 4×A100 GPUs, Qwen3-4B, $G = 8$.

Quantity	GRPO	BICC-GRPO	Ratio
Throughput (tokens/s)	3,231	3,074	0.95×
Peak VRAM per device (GB)	54.6	61.2	1.12×
Wall-clock per step (s)	41.2	43.3	1.05×
GPU-hours to 3k steps	137.3	144.3	1.05×

Scaling to Larger Models. The rollout-to-advantage time ratio is governed by the autoregressive decoding bottleneck and is independent of model scale at the architecture level, so the relative overhead remains comparable when scaling to larger models. We verify this empirically in Appendix D.4.

RCC Overhead. RCC requires $O(G)$ additional operations per query, computing $\delta_i = \log \pi_\theta(o_i|q) - \log \pi_{\theta_{\text{old}}}(o_i|q)$ and the sample covariance $\widehat{\text{Cov}}(R, \delta)$, where $\log \pi_\theta(o_i|q)$ comes from one forward pass of the unaugmented sequence, whose cost is included in the advantage forward pass of Table 10. The total cost is dominated by G subtractions, multiplications, and summations, which is negligible compared to forward and backward passes.

Gradient Variance Measurement. The gradient variance reductions reported in Section 6 and Figure 3b are computed as follows. For each training batch, we collect the per-sample policy gradient vectors $\{\nabla_\theta L_i\}_{i=1}^B$ where B is the batch size and L_i is the per-sample contribution to the clipped surrogate objective. The empirical gradient variance is the trace of the sample covariance matrix:

$$\widehat{\text{Var}}(\nabla_\theta L) = \frac{1}{B-1} \sum_{i=1}^{B} \left\| \nabla_\theta L_i - \overline{\nabla_\theta L} \right\|_2^2, \tag{48}$$

where $\overline{\nabla_\theta L} = \frac{1}{B} \sum_i \nabla_\theta L_i$. Reported variance reduction percentages are averaged over the final 500 training steps, comparing the BICC-GRPO+RCC variant against GRPO under matched batch and seed configurations, and the same reference applies to Table 3(c) and Table 5.

Inference. At inference time, BICC and RCC incur zero overhead since the bilateral context is used only during training. The trained model generates from the original prompt q alone, identical to standard GRPO.

Edge Cases. When either partition is empty ($G^+ = 0$ or $G^- = 0$), bilateral conditioning cannot be applied and BICC falls back to standard GRPO. With $G = 8$, degenerate groups occur in approximately 12% of queries for Qwen3-4B and 18% for Phi-4-mini, concentrated on the easiest and hardest problems, and Table 12 records the rate by training stage. The composition shifts from all-incorrect toward all-correct groups as accuracy rises while the total stays near stationary, so the fallback stays a minority of groups in each training stage. RCC is inert on such a group by construction: with identical rewards every r_i equals $\bar{R}$, the sample covariance of Eq. (19) vanishes for any δ, and Eq. (20) assigns zero advantage, exactly as standard GRPO does at zero reward variance. The correction scales with the reward spread and vanishes continuously, so the transition into and out of a degenerate group keeps the objective continuous. The fallback also composes with dynamic sampling filters such as the one in DAPO (Yu et al., 2025).

Table 12: Share of degenerate groups (%) by training stage, $G = 8$.

Model	Group	0–0.5k	1k–1.5k	2.5k–3k
Qwen3-4B	all incorrect	9.8	6.4	4.1
Qwen3-4B	all correct	3.1	5.2	7.6
Phi-4-mini	all incorrect	16.2	12.1	9.3
Phi-4-mini	all correct	2.4	5.5	8.4

D Supplementary Experiments

D.1 RCC Ablation

Table 13 separates the contributions of BICC and RCC on Math500 with G=8, complementing the Pass@k analysis in Figure 3b.

Table 13: Ablation of BICC and RCC on Math500 (%).

Configuration	Qwen3-4B	Phi-4-mini
GRPO (baseline)	91.4	76.2
BICC-GRPO	92.2	78.1
BICC-GRPO + RCC	92.6	78.8

BICC contributes the primary accuracy gain (+0.8% Qwen, +1.9% Phi). RCC provides additional improvement (+0.4% Qwen, +0.7% Phi), and the combined variant has 31 to 37% lower gradient variance than GRPO in Figure 3b.

D.2 Reward-Confidence Correlation

Table 14 provides numerical values for the $\mathrm{Cov}(R, \delta)$ evolution visualized in Figure 2.

Table 14: $\mathrm{Cov}(R, \delta)$ across training stages.

Model	0–0.5k	0.5k–1k	1k–1.5k	1.5k–2k	2k–2.5k	2.5k–3k
Qwen3-4B	0.015	0.029	0.043	0.054	0.063	0.066
Phi-4-mini	0.037	0.062	0.090	0.113	0.131	0.138

Phi-4-mini exhibits stronger reward-confidence correlation throughout training (0.138 vs 0.066 at convergence), explaining the larger RCC benefit observed for this model.

D.3 Compute-Matched Baseline Construction

The compute-matched row in Table 2 (marked with †) controls for the possibility that BICC's gains stem from the additional FLOPs consumed by extended-context advantage computation instead of the cross-partition contrastive signal.

FLOPs Equalization. BICC-GRPO's training step consumes approximately 5% more total FLOPs than standard GRPO at $G = 8$ with 40% context allocation (Table 10). The overhead concentrates in the advantage computation phase due to extended-context forward passes, while rollout generation FLOPs are unchanged. To approximate this overhead with a standard GRPO baseline of equivalent total compute, we increase the rollout group size from $G = 8$ to $G = 9$, which raises both rollout FLOPs and advantage-computation FLOPs by a factor of $9/8 = 1.125$. Weighted by the 70%/15%/15% phase decomposition in Table 10, this yields a total FLOPs increase of $0.125 \cdot 1.0 \approx 12.5\%$, which exceeds BICC's 5% overhead. We thus interpret the $G = 9$ baseline as a strict upper bound on what GRPO can achieve under matched or higher compute, providing a conservative comparison.

Full Results. Table 15 reports the compute-matched comparison on the four benchmarks, complementing the † row in the main table.

Table 15: Compute-matched comparison: standard GRPO with enlarged rollout budget vs. BICC-GRPO.

	Qwen3-4B				Phi-4-mini			
Method	Math500	AMC 23	AIME 24	AIME 25	Math500	AMC 23	AIME 24	AIME 25
GRPO ($G = 8$)	91.4	72.8	54.0	46.4	76.2	42.5	10.4	4.5
GRPO ($G = 9$, FLOPs-matched)	91.5	73.0	54.1	46.5	76.4	42.7	10.5	4.5
BICC-GRPO ($G = 8$)	**92.2**	**74.2**	**54.6**	**47.1**	**78.1**	**44.3**	**11.2**	**5.1**
Δ (compute-matched vs. GRPO)	+0.1	+0.2	+0.1	+0.1	+0.2	+0.2	+0.1	+0.0
Δ (BICC vs. GRPO)	+0.8	+1.4	+0.6	+0.7	+1.9	+1.8	+0.8	+0.6

The compute-matched row shows that allocating equivalent FLOPs to standard GRPO yields improvements of at most 0.2 percentage points on the four benchmarks of both models, while BICC achieves 0.6 to 1.9 percentage points under the same compute budget. On the seven benchmarks with a nonzero compute-matched gain, the BICC gain is 6 to 9.5 times larger, which indicates that cross-partition contrastive signal, not additional compute, drives the improvement.

D.4 Other Scales and Model Families

We evaluate whether BICC's improvements transfer to larger model scales by training on Qwen3-14B, an instruction-following model 3.5× larger than the Qwen3-4B used in the main experiments, and to a third model family by training Llama-3.1-8B-Instruct under the protocol of Table 6 with 32 evaluation seeds (Table 17).

Table 16: Pass@1 accuracy (%) on Qwen3-14B, $G = 8$, averaged over 32 evaluation seeds.

Method	Math500	AMC 23	AIME 24	AIME 25
GRPO	93.8	78.5	61.5	52.8
BICC-GRPO	**94.3**	**79.4**	**62.0**	**53.4**
Δ	+0.5	+0.9	+0.5	+0.6

BICC improves the four benchmarks at the 14B scale. The absolute gains (0.5 to 0.9 percentage points) are smaller than those observed on Qwen3-4B (0.6 to 1.4) and Phi-4-mini (0.6 to 1.9), in line with the finding of Section 6.2 that weaker models benefit more from explicit contrastive signals. The gains persist at scale, indicating that the cross-partition contrastive mechanism provides value even when base capability is high.

Table 17: Pass@1 accuracy (%) on Llama-3.1-8B-Instruct, $G = 8$, mean $\pm$ standard deviation over 32 seeds.

Method	Math500	AMC 23	AIME 24
GRPO	52.8 $\pm$ 0.44	27.5 $\pm$ 0.61	6.7 $\pm$ 0.79
BICC-GRPO	**54.9** $\pm$ 0.42	**29.6** $\pm$ 0.57	**7.4** $\pm$ 0.76
Δ	+2.1	+2.1	+0.7

The gains on Llama-3.1-8B-Instruct exceed those on Qwen3-4B and Phi-4-mini, following the trend of Section 6 and Appendix E.3 that models with weaker initial capability extract more from the explicit contrastive signal.

Memory and Time Overhead at 14B. On 4×A100 GPUs, BICC's overhead at the 14B scale remains within the same range observed at 4B: total step time 1.06× standard GRPO, peak memory 1.13×. The relative overhead is governed by the rollout-to-advantage-computation ratio, which is approximately scale-invariant for transformer models in the regime tested.

D.5 Cross-Domain Transfer Evaluation

To test whether the cross-partition contrastive mechanism transfers beyond mathematical reasoning, we evaluate BICC-GRPO on logical reasoning and code generation benchmarks. The model is trained exclusively on DAPO-Math-17k (identical to the main experiments) and evaluated zero-shot on the logical reasoning datasets, providing a cross-domain transfer test instead of an in-distribution evaluation.

Datasets. **LogiQA 2.0** contains multi-choice logical reasoning problems requiring structured deduction over natural-language premises. **ZebraLogic** consists of constraint satisfaction puzzles where binary correctness is determined by whether all constraints are satisfied in the model's solution. **HMMT 2025** (Harvard-MIT Mathematics Tournament) provides additional held-out mathematical evaluation distinct from the AIME competition format. **HumanEval** and **MBPP** measure code generation by the unit-test pass rate of the generated program.

Table 18: Cross-domain evaluation: zero-shot transfer from math-trained BICC-GRPO to logical reasoning and code generation. Pass@1 accuracy (%) on Qwen3-4B trained on DAPO-Math-17k only, $G = 8$, 32 seeds.

	In-domain (Math)				Held-out	Logic		Code	
Method	Math500	AMC 23	AIME 24	AIME 25	HMMT 25	LogiQA 2.0	ZebraLogic	HumanEval	MBPP
GRPO	91.4	72.8	54.0	46.4	32.6	63.5	80.4	74.4	66.2
BICC-GRPO	**92.2**	**74.2**	**54.6**	**47.1**	**33.3**	**65.2**	**81.9**	**75.5**	**67.0**
Δ	+0.8	+1.4	+0.6	+0.7	+0.7	+1.7	+1.5	+1.1	+0.8

BICC improves the nine evaluations, including the two logical reasoning and the two code generation benchmarks. The gains on LogiQA 2.0 (+1.7) and ZebraLogic (+1.5) are comparable in magnitude to in-domain mathematical reasoning, indicating that the mechanism of BICC transfers to RLVR settings where binary correctness signals enable partition-based contrastive learning.

Implications. The cross-domain results suggest BICC operates on a structural property of group-based RL (partitioned outcome distribution) and not on domain-specific patterns of mathematical reasoning. The results position BICC as a general-purpose enhancement to GRPO-family methods, not a math-specific technique.

D.6 Sensitivity to the Training Corpus

To separate the mechanism from the composition of DAPO-Math-17k, we retrain Qwen3-4B on the MATH training split (Hendrycks et al., 2021) under hyperparameters identical to the main experiments; the two corpora differ in provenance, scale, and difficulty distribution. Table 19 shows that the absolute accuracies shift with the corpus while the BICC gain changes by at most 0.2 points, so the gain depends on the partition structure that binary rewards induce within each group and not on the corpus.

Table 19: BICC-GRPO under two training corpora, Qwen3-4B, $G = 8$, 32 seeds.

Training corpus	**Method**	Math500	AIME 24
DAPO-Math-17k	GRPO	91.4	54.0
DAPO-Math-17k	BICC-GRPO	**92.2**	**54.6**
MATH train split	GRPO	90.8	53.2
MATH train split	BICC-GRPO	**91.7**	**53.9**

D.7 Ablations on AIME 24

Table 3 reports the ablations on Math500 because its 500 problems resolve differences of 0.2 points, whereas one problem of AIME 24 corresponds to 3.3 points and the seed standard deviation is

0.81 against 0.35 on Math500 (Table 9). Table 20 repeats the three ablations on AIME 24, and the orderings reproduce those of Table 3 in each of the three ablations.

Table 20: Ablations on AIME 24, $G = 8$, Pass@1 (%) over 32 seeds. (a) context allocation ratio in BICC-GRPO; (b) conditioning choice; (c) σ-normalization on top of RCC, Qwen3-4B.

(a) Context allocation ratio.

Ratio	Qwen3	Phi-4
20%	53.9	10.7
40%	**54.6**	**11.2**
60%	54.2	10.9

(b) Conditioning choice.

Type	Qwen3	Phi-4
None	54.0	10.4
Same-class	54.1	10.5
Cross-class	**54.6**	**11.2**

(c) σ-norm in RCC (Qwen3).

Config	Acc.
BICC-GRPO	54.6
+RCC	**54.9**
+RCC+σ	54.2

E Qualitative Analysis and Discussion

E.1 Case Study with Quantitative Conditioning Weights

We illustrate the effect of bilateral conditioning through a representative example. Consider the problem: *Find the smallest positive integer n such that $n^2 + n + 1$ is divisible by 7.* Without bilateral conditioning, the model generates diverse attempts independently. Correct solutions typically identify $n \equiv 2 \pmod 7$ through systematic modular arithmetic (verifying $2^2 + 2 + 1 = 7$), while incorrect attempts test small values without systematic search, misapply the quadratic formula in the modular context, or make arithmetic errors during substitution.

Under BICC, evaluating a correct solution involves observing incorrect attempts as context. This exposure to failure modes strengthens commitment to systematic approaches: the model applies modular arithmetic more reliably instead of trial-and-error. Conversely, when evaluating an incorrect solution, observing correct attempts highlights where reasoning diverges, increasing the gradient penalty for specific error patterns.

Quantitative Evidence. The conditioning weight $w_i = \pi_\theta(o_i|q, \mathcal{O}^{\mp})/\pi_\theta(o_i|q)$ captures this mechanism numerically. Table 4 (Appendix B.4) tracks w_i across training stages, showing $w_i^+ \to 1.14$ for correct samples and $w_i^- \to 0.84$ for incorrect samples by training step 2.5–3k. The divergence from $w_i = 1$ in opposite directions for the two partitions confirms that observing failures amplifies confidence in correct approaches while observing successes attenuates confidence in flawed ones, providing the contrastive gradient signal that drives BICC's improvements.

E.2 Failure Modes

BICC does not universally improve performance. When all sampled outputs are correct ($G^- = 0$) or all incorrect ($G^+ = 0$), bilateral conditioning cannot be applied, and the method falls back to standard GRPO. This occurs more frequently on very easy or very hard problems. Additionally, when incorrect solutions share the same fundamental error (e.g., a shared misreading of the problem), the opposite-partition context provides redundant signal; BICC benefits most when incorrect attempts exhibit diverse failure modes. For problems requiring very long solutions, fewer opposite-partition samples fit within the fixed budget of Section 5.1, which reduces the diversity of failure or success modes that the context represents.

E.3 Why Weaker Models Benefit More

The experimental results show larger BICC improvements on Phi-4-mini than on Qwen3-4B, and three factors may contribute. Weaker models exhibit greater variance in solution quality within each group, which provides a richer contrastive signal, since the bilateral context carries more information when correct and incorrect solutions are more clearly separated. Figure 2 also shows Phi-4-mini develops larger $\mathrm{Cov}(R, \delta)$ during training (0.138 vs 0.066), indicating stronger reward-confidence correlation that makes RCC more impactful. Finally, lower baseline accuracy leaves more room for improvement; the relative improvement may be similar across models even when absolute gains differ.

E.4 Connection to Related Frameworks

BICC shares motivation with self-consistency decoding (Wang et al., 2022), which aggregates multiple samples at inference time. BICC instead leverages sample diversity during training, enabling the model to internalize contrastive patterns without inference-time overhead. The pairwise reformulation (Eq. (11)) also reveals structural similarity to InfoNCE-style objectives; unlike standard contrastive learning where negatives are constructed or retrieved, BICC uses naturally occurring incorrect attempts from the same query, ensuring task-relevant contrast. The observation that weaker base models benefit more suggests BICC implicitly implements a form of curriculum learning, where the model learns to distinguish subtle reasoning errors.

E.5 Limitations and Future Directions

Our theoretical analysis and experiments assume binary rewards, which cover the tasks that RLVR addresses at scale: mathematical reasoning, code generation with unit tests, and constraint satisfaction admit outcome checkers as well, and Appendix D.5 reports transfer to the latter two. Extending to continuous rewards requires a soft partition. The reformulation of Section 4.1 suggests two routes: thresholding the reward at the group median recovers two partitions from continuous rewards, and weighting each pairwise term of Eq. (11) by the reward difference of the pair keeps the contrastive structure with graded membership. Both preserve zero inference overhead, and their empirical study is future work. BICC increases training cost due to longer input sequences; the accuracy-efficiency trade-off is favorable in our experiments, and Appendix D.4 measures a comparable relative overhead at 14B. Training uses mathematical reasoning with verifiable answers and Appendix D.5 evaluates transfer to logic and code, while generalization to open-ended tasks where correctness is not binary remains unexplored. Finally, the effectiveness of BICC depends on reward quality, since noisy or misspecified rewards could produce misleading contrastive signals.

Several directions merit further investigation: learning to selectively weight conditioning samples based on contrastive value, conditioning on solutions from related problems to enable cross-query transfer, and integrating BICC with process reward models for step-level contrastive learning.