ORPG: RECONCILING MULTIPLE REWARD OBJECTIVES THROUGH OBJECTIVE-WISE POLICY GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-reward policy optimization requires a joint update that reflects both the learning signals and the intended relationships among objectives. We introduce Objective-wise Reconciled Policy Gradient (ORPG), which constructs a separate clipped policy objective for each reward and reconciles the resulting gradients into one policy update. For compatible gradients, a cosine-dependent interpolation coordinates their contributions through a partially normalized reference while preserving the norm of their sum. We characterize this update as the unique solution of a spherical directional compromise. For conflicting gradients, projection follows the task's priorities. We evaluate the same compatible rule in helpfulness–safety alignment and correctness–cost optimization for mathematical reasoning. ORPG substantially improves average Useful and Harmless scores over the strongest external baseline on each axis. In mathematics, it achieves the highest average full-budget accuracy and three-budget hypervolume among the compared methods, with more accurate and shorter responses than the initial policy. Component comparisons and training dynamics show the larger contribution of compatible coordination and a complementary benefit from conflict handling. These results support gradient reconciliation for objectives with equal standing and for objectives with an explicit priority.

1 INTRODUCTION

Language models are increasingly expected to satisfy several requirements within the same response. Helpfulness and safety jointly shape the behavior of an assistant (Ouyang et al., 2022; Dai et al., 2023), while mathematical reasoning requires accurate answers at a manageable generation cost (Aggarwal & Welleck, 2025; Liu et al., 2025b). Reinforcement learning provides a way to train for these requirements through separate reward signals. These signals guide the same policy, so learning from them involves deciding how each objective contributes to a shared update. The desired relationship between objectives also depends on the task: helpfulness and safety may have equal standing, whereas reducing generation cost should remain subordinate to correctness. Multi-reward policy optimization therefore requires a way to coordinate learning across objectives while respecting these relationships.

Existing multi-reward optimizers combine objectives at different stages. GRPO-based methods can combine rewards before group-relative advantage estimation (Shao et al., 2024). MO-GRPO and GDPO retain reward-specific statistics when constructing advantages (Ichihara et al., 2025; Liu et al., 2026b), while GD^{2}PO further processes interactions among reward-wise advantages (Liu et al., 2026a). Gradient-based approaches, including GAPO and PAMA, make the shared update direction an explicit design choice (Li et al., 2025; He & Maghsudi, 2025). Separate gradients create a second design problem: how should they form one update? Their directions determine whether contributions reinforce or oppose one another, while their relative norms determine their amplitudes in a direct sum. Conflict projection alone leaves compatible pairs unchanged (Yu et al., 2020), even when one gradient dominates that sum. This motivates coordinating compatible contributions alongside resolving conflicts according to task priorities.

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.

We propose Objective-wise Reconciled Policy Gradient (ORPG), which retains a separate clipped policy objective for each reward and computes its gradient over the same trainable parameters. For a compatible pair, ORPG forms a reference by partially normalizing gradient magnitudes and interpolates between this reference and the original gradient sum. The mixing strength depends on gradient cosine similarity, and a final normalization preserves the sum's norm. For a conflicting pair, ORPG applies symmetric projection to objectives with equal standing, or a one-sided projection that preserves the primary gradient. The same compatible rule therefore serves different conflict priorities. We characterize the compatible update as the unique solution of a directional compromise on a fixed-radius sphere.

We evaluate ORPG on helpfulness–safety alignment and correctness–cost optimization for mathematical reasoning. ORPG improves both Useful and Harmless scores across all three alignment datasets, with average gains of 0.415 and 0.446 over the strongest external baseline on each axis. In mathematics, it achieves the highest average full-budget accuracy and three-budget hypervolume among the compared methods, while improving accuracy and reducing response length relative to the initial policy. Component comparisons in helpfulness–safety identify compatible coordination as the larger source of improvement, with conflict handling providing an additional benefit. Training measurements connect these gains to stronger joint reward learning and the gradient relationships encountered during optimization.

Our contributions are:

- We introduce ORPG, which preserves separate clipped policy objectives and reconciles their full gradients through compatible contribution coordination and task-priority conflict resolution.
- We show that the compatible rule uniquely solves a spherical directional compromise, preserves the original sum norm, and bounds the unit-direction contribution ratio between the original and partially normalized ratios.
- We demonstrate joint helpfulness–safety gains and accuracy-prioritized cost reduction in mathematical reasoning. Component comparisons and training measurements identify the roles of compatible coordination and conflict handling.

2 Related Work

Multi-objective gradient coordination. Gradient-based multi-task methods provide several ways to construct a shared update. GradNorm adapts task weights using gradient magnitudes and relative training rates (Chen et al., 2018). MGDA-based multi-task learning seeks a common descent direction through a combination of objective gradients (Sener & Koltun, 2018), while gradient similarity can regulate auxiliary updates relative to a primary objective (Du et al., 2018). PCGrad projects conflicting gradients (Yu et al., 2020). GradVac adjusts gradient relationships toward target cosine similarities (Wang et al., 2021), and CAGrad controls the worst local objective improvement around the average gradient (Liu et al., 2021). Aligned-MTL constructs the update through an alignment-based transformation of the gradient system (Senushkin et al., 2023). These approaches differ in the quantity they control: task weights, local directional improvement, pairwise relationships, or the conditioning of the gradient system.

Reinforcement learning and multi-reward policy optimization. PPO introduced a clipped policy objective for stable policy updates (Schulman et al., 2017). GRPO estimates advantages from groups of sampled responses and removes the need for a learned value function (Shao et al., 2024). Subsequent multi-objective alignment methods extended the construction of the update. GAPO rescales objective gradients and solves a minimum-norm combination problem (Li et al., 2025), while PAMA combines a modified policy objective with efficient multi-objective weight calculation (He & Maghsudi, 2025). Dynamic reward weighting adapts objective weights during training (Lu et al., 2025). MO-GRPO and GDPO normalize rewards separately before aggregating their advantages (Ichihara et al., 2025; Liu et al., 2026b). Blockwise advantage estimation assigns objective-specific signals to corresponding response blocks (Pavlenko et al., 2026), and GD^{2}PO filters conflicting reward-wise advantages and reweights prompt groups (Liu et al., 2026a). The combination stage determines which interactions the optimizer can act on explicitly. Reward and advantage

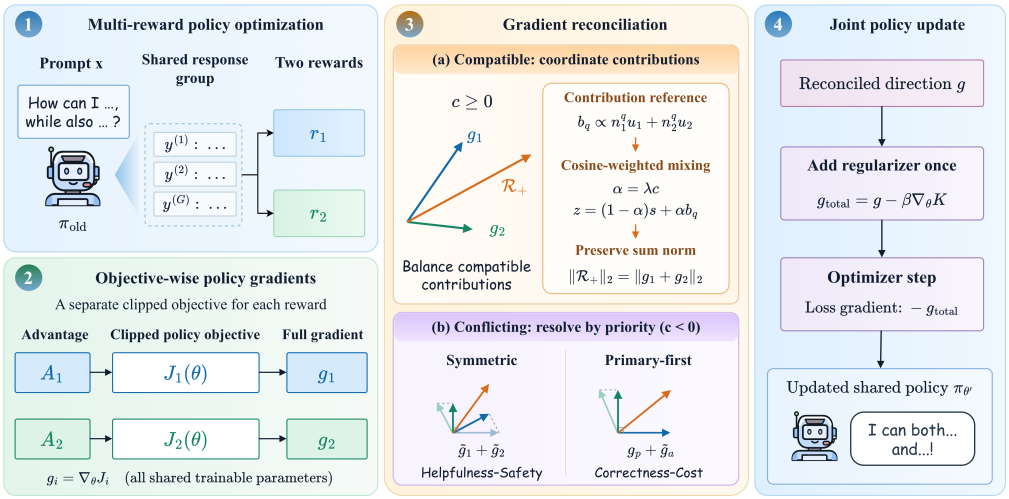

Figure 1: Overview of ORPG. Each reward retains its own clipped policy objective. Reconciliation uses the relationship between their full policy gradients to form a joint update.

methods shape the learning signal before policy differentiation; gradient methods operate on the parameter update induced by that signal.(Li et al., 2026; Zhao et al., 2026) ORPG preserves each reward through a separate clipped policy objective and reconciles the resulting full policy gradients into a joint update.

Length-aware reasoning methods express generation cost through length targets, penalties, or response selection (Aggarwal & Welleck, 2025; Luo et al., 2025a; Yi et al., 2025; Liu et al., 2025a; Shrivastava et al., 2025; Liu et al., 2025b). This setting gives the objectives a primary–secondary relationship: correctness determines answer quality, while length controls the cost of obtaining it.

3 Objective-wise Reconciled Policy Gradient

3.1 Policy optimization setup and separate objectives

Let π_θ be a policy with trainable parameters θ. For each prompt $x \sim \mathcal{D}$, a fixed rollout policy π_{old} samples a group of G responses $y^{(1)}, \ldots, y^{(G)}$. Reward i assigns each response an advantage $A_i^{(j)}$, broadcast over its valid response tokens. GRPO constructs these advantages from within-group reward statistics (Shao et al., 2024); the task-specific constructions used here are given in Section 4 and Appendix B.1.

For token t of response j, define the importance ratio

$$\rho_{j,t}(\theta) = \frac{\pi_\theta(y_t^{(j)} \mid x, y_{<t}^{(j)})}{\pi_{\text{old}}(y_t^{(j)} \mid x, y_{<t}^{(j)})}. \tag{1}$$

PPO-style clipping, also used by GRPO, gives the maximized surrogate integrand (Schulman et al., 2017; Shao et al., 2024)

$$\phi_{\text{clip}}(\rho, A) = \min\{\rho A, \text{clip}(\rho, 1-\epsilon_-, 1+\epsilon_+)A\}. \tag{2}$$

ORPG retains a separate objective for each reward:

$$J_i(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, \{y^{(j)}\}\sim\pi_{\text{old}}} \left[\text{Reduce}_{j,t}\, \phi_{\text{base}}(\rho_{j,t}(\theta), A_i^{(j)}) \right]. \tag{3}$$

Here ϕ_{base} includes the negative-advantage safeguard specified in Appendix B.1. The reduction is a valid-token mean for helpfulness–safety and a sequence mean of token means for mathematics. Separate clipping preserves objective identity through differentiation.

We use ascent notation $g_i = \nabla_\theta J_i$, with each gradient covering all trainable policy parameters. The reconciliation operator $\mathcal{R}$ combines these gradients, followed by one shared regularizer:

$$g_{\text{total}} = \mathcal{R}(g_1, \ldots, g_m) - \beta \nabla_\theta K(\theta). \tag{4}$$

The optimizer uses $-g_{\text{total}}$ as its loss gradient. The framework supports multiple rewards; our implemented and evaluated rule treats two.

3.2 Compatible contributions

For two nonzero gradients, write

$$n_i = \|g_i\|_2, \qquad u_i = g_i/n_i, \qquad c = u_1^\top u_2, \qquad s = g_1 + g_2, \qquad S = \|s\|_2. \tag{5}$$

When $c \geq 0$, both gradients are locally compatible. Their relative norms still determine their amplitudes in $s = n_1 u_1 + n_2 u_2$. ORPG coordinates these amplitudes using a partially normalized reference:

$$v_q = n_1^q u_1 + n_2^q u_2, \qquad b_q = S \frac{v_q}{\|v_q\|_2}, \qquad q \in [0, 1]. \tag{6}$$

The choice $q = 1$ recovers the original sum direction, while $q = 0$ gives equal amplitudes on the unit directions. At $q = \frac{1}{2}$, their ratio becomes $\sqrt{n_1/n_2}$, halfway between equal amplitudes and the original ratio in logarithmic coordinates. This retains information about gradient magnitude while moderating its influence on the joint direction.

Let $\lambda \in [0, 1]$ control the maximum mixing strength and set $\alpha = \lambda c$. The compatible update is

$$z = (1 - \alpha)s + \alpha b_q, \qquad \mathcal{R}_+(g_1, g_2) = S \frac{z}{\|z\|_2}. \tag{7}$$

Cosine similarity controls how strongly the reference contributes. Near orthogonality, the adjustment approaches zero. The final normalization retains the magnitude of the original gradient sum while changing its direction.

Proposition 1 (Spherical directional compromise)**.** *For nonzero g_1, g_2 with $c \geq 0$, $q \in [0, 1]$, and $\lambda \in [0, 1]$, Equation (7) is the unique solution of*

$$\underset{\|g\|_2 = S}{\text{minimize}} \quad (1 - \alpha)\|g - s\|_2^2 + \alpha\|g - b_q\|_2^2. \tag{8}$$

The objective balances proximity to the original sum and to the contribution reference on the same sphere. Expanding the squares reduces the problem to maximizing $g^\top z$ under a norm constraint. Its solution is the normalized vector in Equation (7); a full derivation appears in Appendix A.

Several properties follow directly. The output equals s when $\lambda = 0$, $c = 0$, $q = 1$, the gradients have equal norms, or they point in the same direction. For an unequal-norm pair, the coefficient ratio after mixing lies between the original ratio and its q-power reference. The adjustment therefore changes contributions continuously rather than replacing gradient magnitudes with a binary choice.

3.3 Conflict resolution and priorities

When $d = g_1^\top g_2 < 0$, ORPG uses a conflict rule determined by the task priorities. With symmetric objectives, it applies the two-objective PCGrad projection (Yu et al., 2020):

$$\widetilde{g}_1 = g_1 - \frac{d}{n_2^2} g_2, \qquad \widetilde{g}_2 = g_2 - \frac{d}{n_1^2} g_1, \qquad \mathcal{R}_- = \widetilde{g}_1 + \widetilde{g}_2. \tag{9}$$

Each projected direction removes its component opposing the other objective.

For a primary objective p and a secondary objective a, ORPG preserves g_p and finds the closest secondary direction that does not oppose it:

$$\widetilde{g}_a = \arg\min_h \frac{1}{2}\|h - g_a\|_2^2 \quad \text{subject to} \quad g_p^\top h \geq 0. \tag{10}$$

For a conflicting pair, the closed-form result is

$$\widetilde{g}_a = g_a - \frac{g_p^\top g_a}{\|g_p\|_2^2} g_p, \qquad \mathcal{R}_- = g_p + \widetilde{g}_a. \tag{11}$$

The secondary objective retains its orthogonal component, while the primary direction remains intact. In ascent notation, $g_p^\top \mathcal{R}_- = \|g_p\|_2^2$ for the conflicting pair. This first-order property concerns the reconciled policy direction before shared regularization and the optimizer update.

3.4 Overall update and optimization procedure

At optimization step k, let $\mathcal{B}_k$ be the current minibatch and $g_i^k = \nabla_\theta J_i(\theta; \mathcal{B}_k)|_{\theta=\theta_k}$. The reconciliation rules determine scalar coefficients ω_i^k such that

$$g_{\text{rec}}^k = \mathcal{R}(g_1^k, g_2^k) = \sum_{i=1}^{2} \omega_i^k g_i^k. \tag{12}$$

Holding these coefficients fixed for the current differentiation gives the local surrogate

$$\widetilde{J}_{\text{ORPG}}^k(\theta) = \sum_{i=1}^{2} \text{sg}(\omega_i^k) J_i(\theta; \mathcal{B}_k) - \beta K(\theta; \mathcal{B}_k), \tag{13}$$

where sg denotes stop-gradient. Consequently,

$$\nabla_\theta \widetilde{J}_{\text{ORPG}}^k(\theta)\Big|_{\theta=\theta_k} = g_{\text{rec}}^k - \beta \nabla_\theta K(\theta_k; \mathcal{B}_k) = g_{\text{total}}^k. \tag{14}$$

This representation connects the reconciled direction to the reward-specific objectives. The coefficients are recomputed from the current gradients at every optimization minibatch and remain fixed only for that differentiation. Appendix B.2 gives their closed forms.

Algorithm 1 obtains each full objective gradient separately. Reconciliation then uses three global Gram scalars, n_1^2, n_2^2, d, and $O(P)$ vector operations for P trainable parameters. Appendix B.3 details operations and training costs.

Both symmetric projections use the original gradient pair. When either gradient is zero, the remaining objective passes through unchanged. The shared regularizer is differentiated separately and included once, with $g_K = 0$ when disabled. The optimizer then clips the total loss gradient and applies AdamW. Thus, policy-objective clipping, gradient reconciliation, and final gradient-norm clipping act at distinct stages.

The rollout policy stays fixed within each rollout batch. Both task settings use the same compatible rule; their advantage construction and conflict priority determine how the objectives enter the update.

Algorithm 1 ORPG update

Require: $\pi_\theta, r_1, r_2, q, \lambda, \beta$, task rules
for each rollout batch **do**
  Fix π_{old}; sample response groups
  Evaluate rewards; construct A_1, A_2
  for each optimization minibatch **do**
    Form separate clipped J_1, J_2
    $g_i \leftarrow \nabla J_i$; $g_K \leftarrow \nabla K$
    Compute global n_1^2, n_2^2, d
    if $n_1 n_2 = 0$ **then**
      $g \leftarrow g_1 + g_2$
    else if $d \geq 0$ **then**
      $c \leftarrow d/(n_1 n_2)$; $\alpha \leftarrow \lambda c$
      $g \leftarrow \mathcal{R}_+(g_1, g_2)$ via Eq. (7)
    else if symmetric priority **then**
      $\widetilde{g}_1 \leftarrow g_1 - d g_2/n_2^2$
      $\widetilde{g}_2 \leftarrow g_2 - d g_1/n_1^2$
      $g \leftarrow \widetilde{g}_1 + \widetilde{g}_2$ ▷ Eq. (9)
    else
      Identify primary g_p, secondary g_a
      $\widetilde{g}_a \leftarrow g_a - d g_p/\|g_p\|_2^2$
      $g \leftarrow g_p + \widetilde{g}_a$ ▷ Eq. (11)
    end if
    $h \leftarrow -(g - \beta g_K)$ ▷ Loss gradient
    Clip h to the maximum gradient norm
    Update θ with AdamW using h
  end for
end for

4 Experiments

4.1 Experimental setup

Training and repeated runs. Both settings start from Qwen3-4B-Instruct-2507 (Yang et al., 2025) and use the `verl` framework (Sheng et al., 2024). Evaluation uses 100-step policies. The default Math run accumulates 21.91 training-step hours on eight H200 GPUs, detailed in Appendix B.3. Unless stated otherwise, means and sample standard deviations are computed across three independent training runs and three evaluation runs for base. Appendix B.1 provides optimization settings; Appendix C.1 specifies repeated-run aggregation.

Table 1: Helpfulness–safety results. Each dataset reports Useful (U) and Harmless (H); Avg is the equal-weight average across datasets. Both scores are higher-is-better.

Method	Alpaca		HH-RLHF		PKU-SafeRLHF		Avg	
	U	H	U	H	U	H	U	H
Base	$2.536_{\pm 0.025}$	$2.941_{\pm 0.028}$	$2.855_{\pm 0.002}$	$3.977_{\pm 0.002}$	$4.644_{\pm 0.002}$	$6.404_{\pm 0.004}$	$3.345_{\pm 0.009}$	$4.441_{\pm 0.010}$
GRPO	$5.232_{\pm 0.036}$	$6.212_{\pm 0.080}$	$4.470_{\pm 0.037}$	$6.082_{\pm 0.044}$	$5.620_{\pm 0.011}$	$6.938_{\pm 0.002}$	$5.107_{\pm 0.028}$	$6.411_{\pm 0.038}$
GDPO	$5.335_{\pm 0.010}$	$6.250_{\pm 0.121}$	$4.546_{\pm 0.004}$	$6.116_{\pm 0.019}$	$5.640_{\pm 0.008}$	$6.939_{\pm 0.005}$	$5.174_{\pm 0.005}$	$6.435_{\pm 0.047}$
GD^{2}PO	$5.332_{\pm 0.010}$	$6.297_{\pm 0.033}$	$4.541_{\pm 0.007}$	$6.138_{\pm 0.010}$	$5.631_{\pm 0.001}$	$6.939_{\pm 0.002}$	$5.168_{\pm 0.005}$	$6.458_{\pm 0.013}$
ORPG	$\mathbf{5.943}_{\pm 0.002}$	$\mathbf{7.061}_{\pm 0.002}$	$\mathbf{5.044}_{\pm 0.002}$	$\mathbf{6.679}_{\pm 0.003}$	$\mathbf{5.781}_{\pm 0.0001}$	$\mathbf{6.972}_{\pm 0.0002}$	$\mathbf{5.589}_{\pm 0.0004}$	$\mathbf{6.904}_{\pm 0.001}$

Data, objectives, and benchmarks. For helpfulness–safety, we follow Safe RLHF's separation of alignment criteria (Dai et al., 2023), using the Artessay Qwen2.5-7B-SafeRLHF reward and cost models to score helpfulness and harmlessness(Yang et al., 2024; Artessay, n.d.b;n). Training uses Alpaca with disjoint calibration and evaluation subsets (Taori et al., 2023); evaluation covers Alpaca, HH-RLHF, and PKU-SafeRLHF (Bai et al., 2022; Ji et al., 2025). Group-centered advantages share a scale, and conflict resolution is symmetric. For correctness–cost, training uses DeepScaleR preview prompts (Luo et al., 2025b); evaluation covers AIME-24, AMC-22-23, MATH, Minerva-Math, and OlympiadBench (Hugging Face H4, n.d.; AI-MO, n.d.; Hendrycks et al., 2021; Lewkowycz et al., 2022; He et al., 2024). The rewards are binary correctness and an indicator of length at most $\tau = 4000$. ORPG and all external training baselines use these same reward definitions and length threshold. Correctness uses group-relative advantages and receives conflict priority; length advantages are centered within the correct subset and zero elsewhere. Appendices B.1 and C.1 detail objective construction, dataset sizes, and prompts.

Baselines. We compare ORPG with the initial model, GRPO, GDPO, and the hard variant of GD^{2}PO (Shao et al., 2024; Liu et al., 2026b;a). The initial policy anchors changes in task quality and generation cost. GRPO combines rewards before constructing its group-relative update. GDPO separately normalizes reward-wise advantages before aggregation, while GD^{2}PO filters conflicting reward-wise advantages and reweights prompt groups. These comparisons distinguish coordination at the learning-signal level from reconciliation of separate policy gradients. Section 4.3 evaluates the reconciliation components and alternative gradient combination rules under the objective-wise formulation.

Evaluation metrics. For helpfulness–safety, each evaluation run generates one response for every prompt in each complete set. We report mean Useful and Harmless scores. For mathematics, accuracy estimates pass@1 from four responses per problem. Table 2 reports our primary comparison: accuracy and mean length at the 8192-token budget. Table 3 reports hypervolume (HV), which summarizes the accuracy–cost trade-off using 2048-, 4096-, and 8192-token measurements. Shorter-budget responses are exact prefixes of the same generations. For each dataset, HV is the union area of rectangles from $(0, 0)$ to accuracy–efficiency points $(a, 1 - \ell/8192)$, with efficiency clipped to $[0, 1]$. Avg weights datasets equally after computing their metrics. Appendix C.2 gives the scoring protocol, formula, and budget-specific values.

4.2 Results

Table 1 shows that ORPG achieves the highest Useful and Harmless scores on all three evaluation sets. Its average Useful score of 5.589 exceeds GDPO by 0.415, while its average Harmless score of 6.904 exceeds GD^{2}PO by 0.446. Both scores improve within each dataset, covering general instructions and the two safety-oriented evaluation sets.

Table 2 shows the accuracy-priority outcome in mathematical reasoning. ORPG achieves 66.7% average accuracy at the 8192-token budget, improving on Base by 1.14 percentage points while using 611 fewer tokens per response (19.3%). It exceeds all three external training baselines in accuracy on every dataset. Relative to Base, four datasets improve and AMC-22-23 retains the same mean accuracy. The external baselines produce shorter responses than ORPG but reduce accuracy

Table 2: Mathematical accuracy (Acc, %) and mean response length (Len, tokens) at the 8192-token budget. Each dataset is followed by the equal-weight Avg. A24: AIME-24; AMC: AMC-22-23; Min.: Minerva-Math; Oly.: OlympiadBench.

Method	A24		AMC		MATH		Min.		Oly.		Avg	
	Acc (%)	Len	Acc (%)	Len	Acc (%)	Len	Acc (%)	Len	Acc (%)	Len	Acc (%)	Len
Base	$53.9_{\pm 3.4}$	$5698_{\pm 60}$	$\mathbf{82.4}_{\pm 0.7}$	$3351_{\pm 86}$	$87.6_{\pm 0.1}$	$1462_{\pm 6}$	$37.3_{\pm 0.3}$	$1495_{\pm 22}$	$66.4_{\pm 0.2}$	$3821_{\pm 9}$	$65.5_{\pm 0.7}$	$3165_{\pm 12}$
GRPO	$42.2_{\pm 1.0}$	$2541_{\pm 135}$	$75.5_{\pm 2.3}$	$1708_{\pm 58}$	$86.0_{\pm 0.2}$	$939_{\pm 40}$	$37.9_{\pm 0.6}$	$964_{\pm 91}$	$61.8_{\pm 0.3}$	$1688_{\pm 71}$	$60.7_{\pm 0.5}$	$1568_{\pm 66}$
GDPO	$38.6_{\pm 1.3}$	$\mathbf{2255}_{\pm 105}$	$74.7_{\pm 0.6}$	$\mathbf{1540}_{\pm 26}$	$85.5_{\pm 0.1}$	$\mathbf{833}_{\pm 12}$	$37.3_{\pm 0.6}$	$833_{\pm 25}$	$60.6_{\pm 0.6}$	$\mathbf{1501}_{\pm 46}$	$59.3_{\pm 0.1}$	$\mathbf{1392}_{\pm 38}$
GD^{2}PO	$41.9_{\pm 5.4}$	$2470_{\pm 90}$	$75.1_{\pm 1.5}$	$1583_{\pm 51}$	$85.4_{\pm 0.1}$	$846_{\pm 12}$	$37.9_{\pm 1.0}$	$\mathbf{826}_{\pm 34}$	$60.9_{\pm 0.3}$	$1553_{\pm 29}$	$60.3_{\pm 0.8}$	$1456_{\pm 29}$
ORPG	$\mathbf{56.4}_{\pm 1.9}$	$4613_{\pm 94}$	$\mathbf{82.4}_{\pm 0.6}$	$2748_{\pm 39}$	$\mathbf{88.0}_{\pm 0.1}$	$1229_{\pm 3}$	$\mathbf{39.3}_{\pm 0.5}$	$1287_{\pm 27}$	$\mathbf{67.2}_{\pm 0.6}$	$2893_{\pm 23}$	$\mathbf{66.7}_{\pm 0.3}$	$2554_{\pm 22}$

Table 3: Mathematical hypervolume (HV, higher is better) over the 2048-, 4096-, and 8192-token budgets. Avg gives equal weight to each dataset.

Method	A24	AMC	MATH	Min.	Oly.	Avg
Base	$0.280_{\pm 0.015}$	$0.608_{\pm 0.006}$	$0.768_{\pm 0.001}$	$0.325_{\pm 0.002}$	$0.485_{\pm 0.0005}$	$0.493_{\pm 0.004}$
GRPO	$0.311_{\pm 0.011}$	$0.620_{\pm 0.014}$	$0.769_{\pm 0.003}$	$0.336_{\pm 0.003}$	$0.507_{\pm 0.007}$	$0.509_{\pm 0.003}$
GDPO	$0.295_{\pm 0.011}$	$0.625_{\pm 0.005}$	$0.774_{\pm 0.001}$	$0.337_{\pm 0.006}$	$0.507_{\pm 0.006}$	$0.507_{\pm 0.002}$
GD^{2}PO	$0.315_{\pm 0.040}$	$0.627_{\pm 0.014}$	$0.773_{\pm 0.001}$	$0.342_{\pm 0.009}$	$0.508_{\pm 0.002}$	$0.513_{\pm 0.005}$
ORPG	$\mathbf{0.339}_{\pm 0.010}$	$\mathbf{0.637}_{\pm 0.004}$	$\mathbf{0.777}_{\pm 0.0005}$	$\mathbf{0.344}_{\pm 0.005}$	$\mathbf{0.516}_{\pm 0.004}$	$\mathbf{0.523}_{\pm 0.002}$

relative to Base: their average accuracies range from 59.35% to 60.69%. ORPG obtains its cost reduction while improving the primary correctness objective.

Table 3 evaluates the joint accuracy–cost outcome across the three budgets. ORPG reaches an average HV of 0.523, compared with 0.493 for Base and 0.513 for GD^{2}PO, the strongest external baseline on this metric. ORPG's leading full-budget accuracy and aggregate HV show improved correctness and joint accuracy–cost performance. Appendix C.2 provides the budget-specific measurements used to compute HV.

4.3 Component contributions

All variants in Table 4 retain the same objective-wise policy structure. Without either component uses the direct gradient sum (Sum). Without compatible coordination applies conflict projection and directly sums compatible gradients (PCGrad). Without conflict resolution retains compatible coordination and directly sums opposing gradients. CAGrad and Aligned-MTL replace the reconciliation operator with their respective joint-gradient rules. Appendix D.1 gives the formulas.

Table 4 shows that ORPG achieves the highest Useful and Harmless scores among the objective-wise rules. Its improvement over the version without either component establishes the benefit of coordinating the gradients after preserving separate policy objectives. Compatible coordination provides the larger component gain: removing it reduces Useful by 0.378 and Harmless by 0.193. Removing compatible coordination gives results close to removing both components, while the compatible-only variant approaches the full method. The distinction between these updates is how they combine compatible gradients, connecting the largest gain to the central design choice in ORPG.

Conflict resolution further improves both scores. ORPG also exceeds CAGrad and Aligned-MTL on both axes. Section 4.4 examines the learning dynamics. In mathematics, the full rule achieves the highest 8192-budget accuracy among the four component versions, exceeding the version without conflict resolution by 1.20 percentage points. Appendix D.2 reports the accuracy–cost comparison.

All five configurations per setting outperform the external training baselines on both HS scores, full-budget mathematical accuracy, and HV. The default leads in HS scores and full-budget accuracy; $q = 0.75$ achieves higher mathematical HV. Appendix D.3 reports the individual q and λ scans.

Table 4: Reconciliation components and alternative gradient rules on helpfulness–safety. All rows retain separate policy objectives. Scores are averaged over the same three datasets as Table 1.

Update	Useful ↑	Harmless ↑
ORPG	**5.589**$_{\pm 0.0004}$	**6.904**$_{\pm 0.001}$
Without compatible coordination	5.212$_{\pm 0.062}$	6.710$_{\pm 0.032}$
Without conflict resolution	5.563$_{\pm 0.0006}$	6.859$_{\pm 0.001}$
Without either component	5.193$_{\pm 0.073}$	6.695$_{\pm 0.085}$
CAGrad	5.183$_{\pm 0.002}$	6.626$_{\pm 0.002}$
Aligned-MTL	5.065$_{\pm 0.002}$	6.501$_{\pm 0.005}$

4.4 Training dynamics

ORPG learns stronger usefulness and harmlessness together during training. We compare calibrated training rewards against external policy optimizers and objective-wise gradient rules in Figures 2 and 3. Appendix E.2 gives the run-level statistics and advantage measurements.

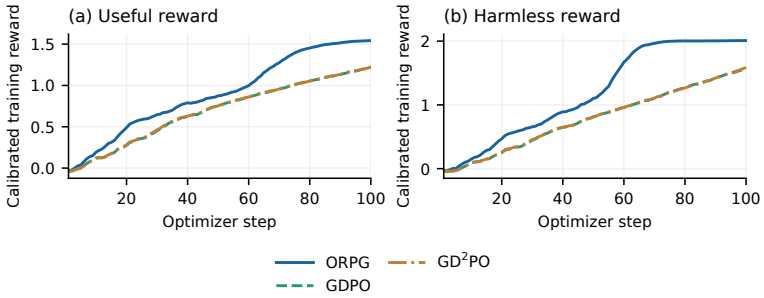

Figure 2: Training rewards against external baselines. (a) Useful and (b) Harmless. A trailing five-step mean is applied to each curve. These calibrated training rewards differ from the raw evaluation scores in Table 1.

Stronger joint learning than external baselines. Figures 2(a–b) show that ORPG develops an advantage on both rewards and extends it through the middle and later stages. The separation is especially visible around steps 60–80: Useful continues to rise while Harmless reaches a higher level. Over the final 20 steps, ORPG averages 1.522 Useful and 2.005 Harmless, compared with 1.157 and 1.458 for GDPO, and 1.158 and 1.462 for GD^{2}PO. The simultaneous gains connect the stronger held-out scores to improved learning of both training objectives.

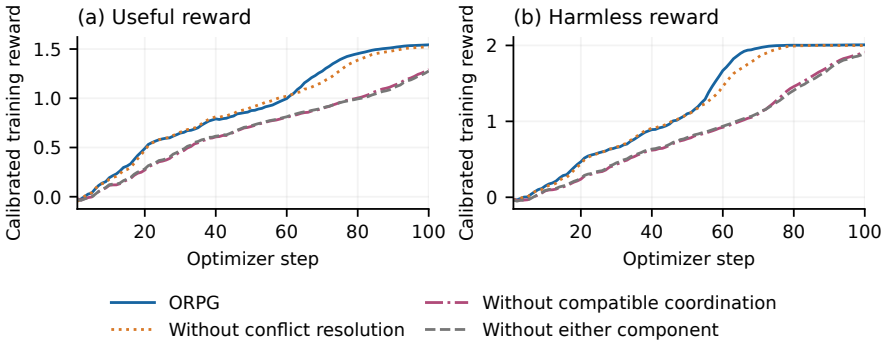

Figure 3: Training rewards under objective-wise gradient rules. (a) Useful and (b) Harmless. Colors and line styles identify the same versions throughout. Smoothing matches Figure 2.

Compatible coordination provides the main reward gain. Figures 3(a–b) separate the four component versions. The two retaining compatible coordination develop substantially higher rewards in the middle and later stages, while removing this component gives a trajectory close to removing both. Over the final 20 steps, ORPG reaches 1.522 Useful and 2.005 Harmless, compared with 1.492 and 1.999 without conflict resolution. Both exceed the versions without compatible coordination and without either component. Together with Table 4, these trajectories identify the main gain from compatible coordination and the additional improvement from the complete rule.

Learning improves in a predominantly compatible regime. Figure 4(a) shows positive step-average gradient cosine through most of training for all four versions. The reward gains from compatible coordination therefore develop largely in a regime where conflict projection leaves the gradient sum unchanged. This connects the training behavior to the motivation for coordinating contributions even when gradients are locally compatible. The versions without compatible coordination and without either component record zero conflicts, whereas ORPG and the version without conflict resolution encounter conflicts late in training; ORPG's average projection rate is 6.5%. Appendix E.1 gives conflict and projection trajectories and stage summaries.

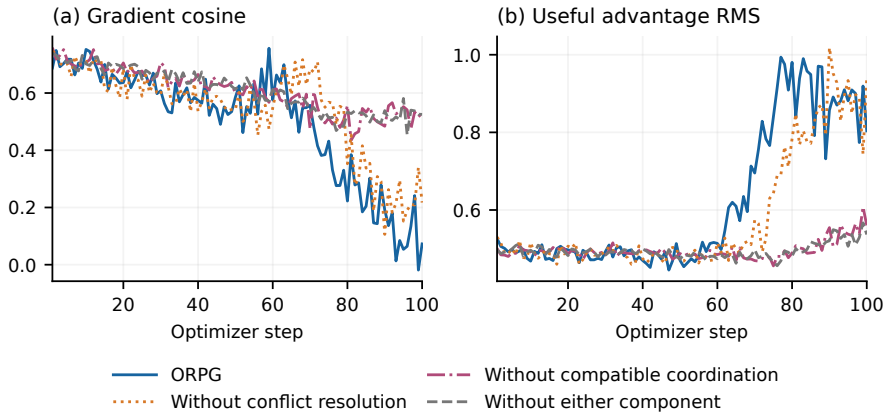

Figure 4: Training signals for the four component versions. (a) Gradient cosine and (b) Useful advantage RMS, shown as unsmoothed step aggregates. Colors and line styles match Figure 3.

A stronger Useful signal accompanies reward improvement. Figure 4(b) shows that the versions retaining compatible coordination sustain a stronger Useful advantage signal later in training. Over the final 20 steps, Useful RMS reaches 0.887 for ORPG and 0.860 without conflict resolution, compared with 0.528 without compatible coordination and 0.521 without either component. All four use the same group-centering and shared-scale advantage construction. The stronger signal accompanies the higher Useful reward in Figure 3(a), indicating effective optimization of usefulness alongside the higher harmlessness reward.

5 Conclusion

We propose ORPG, a gradient-reconciliation method for multi-reward policy optimization that improves joint performance across reward objectives. Its compatible branch coordinates relative contributions while preserving the norm of the gradient sum, and its conflict branch follows task priorities. ORPG improves usefulness and harmlessness jointly and achieves the highest average full-budget accuracy and three-budget hypervolume in mathematical reasoning among the compared methods. Component comparisons and training dynamics in helpfulness–safety identify compatible coordination as the larger source of improvement, with conflict handling adding a complementary benefit. Across the two settings, the results support coordinating objectives according to both their local gradient relationships and their task-level priorities. In summary, ORPG opens a research direction for multi-objective policy optimization through objective-wise gradient reconciliation.

AI USE STATEMENT

In this work, we used generative AI writing assistants for polishing the English of manuscript drafts. We have not used generative AI tools for generating research ideas, hypotheses, or experimental designs, and no reported number was produced by a generative AI tool: all results come from the deterministic training and evaluation pipelines described in the paper. We have reviewed all AI-assisted work: AI-assisted code was checked against the written specification and is covered by the property tests and the coordinate-reflection audit, and AI-polished text was revised and approved by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the update, with theoretical properties derived in Appendix A. Appendices B.1 and B.2 describe objective construction, optimization settings, and reconciliation coefficients. Section 4 and Appendix C.1 provide the task and evaluation protocols, while Appendix C.2 defines the accuracy–cost metrics. Appendices E.1 and E.2 detail the training-gradient and reward measurements.

REFERENCES

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025. URL https://arxiv.org/abs/2503.04697.

AI-MO. AIMO Validation: AMC, n.d. URL https://huggingface.co/datasets/AI-MO/aimo-validation-amc. Dataset release; accessed September 15, 2026.

Artessay. Qwen2.5-7B-SafeRLHF-CM, n.d.a. URL https://www.modelscope.cn/models/Artessay/Qwen2.5-7B-SafeRLHF-CM. Model card; accessed September 15, 2026.

Artessay. Qwen2.5-7B-SafeRLHF-RM, n.d.b. URL https://www.modelscope.cn/models/Artessay/Qwen2.5-7B-SafeRLHF-RM. Model card; accessed September 15, 2026.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023. URL https://arxiv.org/abs/2310.12773.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity, 2018. URL https://arxiv.org/abs/1812.02224.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.

3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Qiang He and Setareh Maghsudi. Pareto Multi-Objective Alignment for Language Models, 2025. URL https://arxiv.org/abs/2508.07768.

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, 2021. URL https://arxiv.org/abs/2103.03874.

Hugging Face H4. AIME 2024 Dataset, n.d. URL https://huggingface.co/datasets/HuggingFaceH4/aime_2024. Dataset release; accessed September 15, 2026.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Mitsuki Sakamoto, Ryota Mitsuhashi, and Eiji Uchibe. MO-GRPO: Mitigating Reward Hacking of Group Relative Policy Optimization on Multi-Objective Problems, 2025. URL https://arxiv.org/abs/2509.22047.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1544. URL https://aclanthology.org/2025.acl-long.1544/.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11214–11232, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.549. URL https://aclanthology.org/2025.acl-long.549/.

Ze-Ping Li, Hongru Wang, Yiwen Zhao, Guanhua Chen, Yixia Li, Keyang Chen, Yixin Cao, Guangnan Ye, Hongfeng Chai, and Zhen-Fei Yin. Rethinking the role of entropy in optimizing tool-use behaviors for large language model agents, 2026. URL https://api.semanticscholar.org/CorpusID:285271471.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-Averse Gradient Descent for Multi-task learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9d27fdf2477ffbff837d73ef7ae23db9-Paper.pdf.

Haotian Liu, Yihao Liu, Jingwei Ni, Siyuan Huang, Xinpeng Liu, Pengyu Cheng, Jiajun Song, Ruijin Ding, Junfeng Li, Zhechao Yu, Mengyu Zhou, Hongteng Xu, Xiaoxi Jiang, and Guanjun Jiang. GD^{2}PO: Mitigating Multi-Reward Conflicts via Group-Dynamic reward-Decoupled Policy Optimization, 2026a. URL https://arxiv.org/abs/2606.16771.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. DLER: Doing length penalty right – incentivizing more intelligence per token via reinforcement learning. *arXiv preprint arXiv:2510.15110*, 2025a. URL https://arxiv.org/abs/2510.15110.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-Decoupled Normalization Policy Optimization for Multi-reward RL Optimization, 2026b. URL https://arxiv.org/abs/2601.05242.

Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. Learn to reason efficiently with adaptive length-based reward shaping. *arXiv preprint arXiv:2505.15612*, 2025b. URL https://arxiv.org/abs/2505.15612.

Yining Lu, Zilong Wang, Shiyang Li, Xin Liu, Changlong Yu, Qingyu Yin, Zhan Shi, Zixuan Zhang, and Meng Jiang. Learning to Optimize Multi-Objective Alignment Through Dynamic Reward Weighting, 2025. URL https://arxiv.org/abs/2509.11452.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-Pruner: Length-harmonizing fine-tuning for O1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025a. URL https://arxiv.org/abs/2501.12570.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, et al. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025b. Notion Blog.

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

Kirill Pavlenko, Alexander Golubev, Simon Karasik, and Boris Yangel. Blockwise Advantage Estimation for Multi-Objective RL with Verifiable Rewards, 2026. URL https://arxiv.org/abs/2602.10231.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html.

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent Component Alignment for Multi-Task Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20083–20093, June 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256.

Vaishnavi Shrivastava, Ahmed Awadallah, Vidhisha Balachandran, Shivam Garg, Harkirat Behl, and Dimitris Papailiopoulos. Sample more to think less: Group filtered policy optimization for concise reasoning. *arXiv preprint arXiv:2508.09726*, 2025. URL https://arxiv.org/abs/2508.09726.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient Vaccine: Investigating and Improving Multi-task Optimization in Massively Multilingual Models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf/4958372042631716242a4b3f1a10231548614522.pdf.

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

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhen-Ru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Jingyang Yi, Jiazheng Wang, and Sida Li. ShorterBetter: Guiding reasoning models to find optimal inference length for efficient reasoning. *arXiv preprint arXiv:2504.21370*, 2025. URL https://arxiv.org/abs/2504.21370.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient Surgery for Multi-Task Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

Yiwen Zhao, Zhihao Wen, Yuchen Mao, Mingxuan Jiang, Yihao Hu, Pan Wang, Xin Zhang, and Wei Wu. Towards better agents for multi-turn user interaction: The next user turn is more than context, 2026. URL https://api.semanticscholar.org/CorpusID:291184908.

A Properties of the reconciliation operator

A.1 Proof of Proposition 1

For compatible nonzero gradients, $S > 0$ and $v_q \neq 0$. Moreover,

$$s^\top v_q = n_1^{q+1} + n_2^{q+1} + c(n_1 n_2^q + n_2 n_1^q) > 0.$$

Thus $s^\top b_q > 0$, so $z = (1-\alpha)s + \alpha b_q$ is nonzero for all $\alpha \in [0,1]$. Since $\|s\| = \|b_q\| = \|g\| = S$, the objective in Equation (8) is

$$2S^2 - 2g^\top\big((1-\alpha)s + \alpha b_q\big) = 2S^2 - 2g^\top z.$$

Cauchy–Schwarz gives $g^\top z \leq S\|z\|$, with equality only at $g = Sz/\|z\|$. This proves the unique minimizer.

A.2 Contribution ratios and identity cases

Before the final normalization, $z = a_1 u_1 + a_2 u_2$, where

$$a_i = (1-\alpha)n_i + \frac{\alpha S}{\|v_q\|} n_i^q.$$

Assume $r = n_1/n_2 \geq 1$. Let $w = (1-\alpha)n_2$ and $t = \alpha S n_2^q/\|v_q\|$. Then

$$\frac{a_1}{a_2} = \frac{wr + tr^q}{w + t} \in [r^q, r].$$

The common final normalization does not change this ratio. At $q = \frac{1}{2}$, the reference log-ratio is half the original log-ratio. If $\alpha = 0$ or $q = 1$, Equation (7) returns s. Equal norms make v_q proportional to s, as does $c = 1$, so these cases also return s. For $q > 0$, the contribution of a vanishing objective tends to zero; the implementation passes through the other gradient when one objective is inactive.

These properties describe the reconciled policy gradient. The shared regularizer and the optimizer act after reconciliation, as specified in Equation (4).

A.3 Priority projection

The feasible set $\{h : g_p^\top h \geq 0\}$ is a closed half-space. If $g_p^\top g_a < 0$, its Euclidean projection is $\widetilde{g}_a = g_a - (g_p^\top g_a) g_p/\|g_p\|^2$. It follows that $g_p^\top \widetilde{g}_a = 0$, so the joint direction has primary directional derivative $\|g_p\|^2$. For a compatible pair, the operator instead uses Equation (7).

A.4 Compatible coordination for multiple objectives

The objective-wise construction in Equation (3) permits any number of reward objectives. The compatible reference also admits a direct extension. Let I index the nonzero gradients and suppose $g_i^\top g_j \geq 0$ for all $i, j \in I$. For $q \in [0,1]$, define

$$s = \sum_{i \in I} g_i, \qquad S = \|s\|, \qquad v_q = \sum_{i \in I} \|g_i\|^{q-1} g_i, \qquad b_q = S\frac{v_q}{\|v_q\|}. \tag{15}$$

For nonempty I, both S and $\|v_q\|$ are positive: the squared norms contain positive diagonal terms and nonnegative cross terms. Moreover,

$$s^\top v_q = \sum_{i \in I} \|g_i\|^{q+1} + \sum_{\substack{i,j \in I \\ i \neq j}} \|g_j\|^{q-1} g_i^\top g_j > 0. \tag{16}$$

Thus $s^\top b_q > 0$. Given any mixing weight $\alpha \in [0,1]$, the vector $z = (1-\alpha)s + \alpha b_q$ is nonzero, and the same spherical compromise has the unique solution

$$\arg\min_{\|g\|=S} \big[(1-\alpha)\|g - s\|^2 + \alpha\|g - b_q\|^2\big] = S\frac{z}{\|z\|}. \tag{17}$$

Expanding the objective gives a constant minus $2g^\top z$, so the result follows by maximizing the inner product on the sphere. This extends Proposition 1 to the reference in Equation (15). Zero gradients are excluded before evaluating the norm powers; if all gradients are zero, the update is zero.

This construction specifies compatible coordination given a mixing weight. A complete rule for multiple objectives additionally requires a choice of α from their joint geometry and a conflict operator for mixed relationships and task priorities. The implemented and evaluated rule in this paper specifies these choices for two objectives.

B Implementation and computational cost

B.1 Objective construction and implementation

In the helpfulness–safety setting, the reward scores are calibrated using fixed statistics estimated from the held-out training-calibration subset. Each objective is centered within a response group. A common scale is then applied to the components, retaining their separate values. No evaluation prompt is part of the calibration subset.

For mathematical reasoning, let C denote the correct responses in a group. The secondary advantage is

$$A_{L,j} = \begin{cases} r_{L,j} - |C|^{-1} \sum_{k \in C} r_{L,k}, & j \in C,\ |C| > 0, \\ 0, & \text{otherwise.} \end{cases}$$

If all responses are incorrect, the secondary objective is inactive. If the length reward is constant on the correct subset, it contributes zero. The correctness advantage retains its primary GRPO normalization. For each response group, correctness scores are centered and divided by their sample standard deviation plus 10^{-6}. In helpfulness–safety, let B_i be the group-centered component broadcast over valid response tokens and M their mask. The common normalization is

$$A_i = \frac{B_i - \mathrm{Mean}_M(B_i)}{\sqrt{\mathrm{Var}_M(B_1 + B_2) + 10^{-8}}}\, M.$$

Here Var_M is the masked variance used by the policy-training implementation. The masked variance applies Bessel's correction over valid response tokens.

The policy-loss adapter differentiates each reward-specific loss over all trainable policy parameters. Distributed reductions produce the global Gram entries. The shared regularization contribution is differentiated separately and included once. The maximized policy surrogate uses asymmetric clipping and a negative-advantage safeguard:

$$\phi_{\mathrm{base}}(\rho, A) = \begin{cases} \max\{\phi_{\mathrm{clip}}(\rho, A), \kappa A\}, & A < 0, \\ \phi_{\mathrm{clip}}(\rho, A), & A \geq 0. \end{cases}$$

Both scenarios use $\kappa = 3$, $\epsilon_- = 0.2$, and $\epsilon_+ = 0.28$. Helpfulness–safety averages over all valid response tokens in the minibatch. Mathematics first averages valid tokens within each response and then averages responses. Exact-zero objective advantages remain zero after normalization, and an inactive objective contributes no gradient.

The mathematical setting uses the MSE log-ratio regularizer

$$K(\theta; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \frac{1}{T_j} \sum_{t=1}^{T_j} \frac{1}{2} \left(\log \pi_\theta(y_t^{(j)} \mid x_j, y_{<t}^{(j)}) - \log \pi_{\mathrm{ref}}(y_t^{(j)} \mid x_j, y_{<t}^{(j)}) \right)^2,$$

where T_j counts valid response tokens and π_{ref} is the frozen initial policy. The reduction is the same sequence mean of token means used for the policy objectives, and $\beta = 0.0005$. Helpfulness–safety disables this regularizer. Both settings use zero entropy coefficient.

Table 5 lists the training settings for the complete method in both scenarios. All policy parameters are trained. The initial policy is Qwen3-4B-Instruct-2507; helpfulness and harmlessness are scored by the Artessay Qwen2.5-7B-SafeRLHF reward and cost models.

Mathematical training uses eight H200 GPUs on one node. LoRA is disabled. Repeated-run aggregation is specified below.

Table 5: Training settings. Token limits refer to training; evaluation budgets are specified separately.

Parameter	Helpfulness–safety	Correctness–cost
Training steps	100	100
Batch / minibatch / rollout group	512 / 128 / 4	512 / 64 / 8
PPO epochs	1	1
Learning rate	2×10^{-6}	10^{-6}
Optimizer	AdamW	AdamW
Schedule / warmup	Constant / none	Constant / none
Adam (β_1, β_2) / ϵ	$(0.9, 0.999)$ / 10^{-8}	$(0.9, 0.999)$ / 10^{-8}
Weight decay	0.01	0.01
Prompt / response limit	512 / 1024	1024 / 8000
Temperature / top-p / top-k	0.7 / 1.0 / -1	1.0 / 1.0 / -1
PPO clip lower / upper	0.2 / 0.28	0.2 / 0.28
Loss reduction	Token mean	Sequence mean of token means
Gradient norm clip	1.0	1.0
KL coefficient / type	0 / disabled	0.0005 / MSE
Entropy coefficient	0	0
Compatible q / λ	0.5 / 0.25	0.5 / 0.25
Preserve sum norm	Yes	Yes
Conflict rule	Symmetric	Correctness priority
Reward weights	1 / 1	1 / 1
Length threshold	—	$\tau = 4000$
Advantage construction	Group centered, shared scale	Primary GRPO; correct-subset length
Precision / sharding	bfloat16 / FSDP	bfloat16 / FSDP

B.2 Reconciliation coefficients and implementation details

Algorithm 1 computes the joint direction directly. Its rules can also be written as the weighted gradient in Equation (12). All quantities below are evaluated at the current minibatch and parameters; their step superscripts are omitted. For nonzero compatible gradients,

$$\omega_i = \frac{S}{\|z\|_2} \left[(1-\alpha) + \frac{\alpha S n_i^{q-1}}{\|v_q\|_2} \right], \qquad i \in \{1, 2\}.$$

For a symmetric conflicting pair with $d = g_1^\top g_2 < 0$,

$$\omega_i = 1 - \frac{d}{n_i^2}, \qquad i \in \{1, 2\}.$$

For primary–secondary conflict resolution,

$$\omega_p = 1 - \frac{d}{\|g_p\|_2^2}, \qquad \omega_a = 1.$$

When either gradient is zero, $\omega_1 = \omega_2 = 1$ gives the direct sum. These coefficients are algebraic expansions of the reconciliation rules. They are recomputed at each optimization minibatch and held fixed in the local surrogate's differentiation. The coefficient ratio ω_1/ω_2 weights the original gradients; the ratio a_1/a_2 in Appendix A weights their unit directions.

B.3 Training cost and reconciliation operations

The default mathematical run uses eight H200 GPUs for 100 optimizer steps. Summing the recorded duration of these steps gives 21.91 hours, or 175.31 GPU-hours, with a mean of 788.9 seconds per step. This measures accumulated training-step time, excluding queueing, intervals between training processes, initialization outside the step timers, discarded progress, and separate validation and evaluation. Table 6 reports the recorded components. Rollout generation averages 211.1 seconds per step and actor updates 474.4 seconds. Component timers may nest, so their entries are not an additive partition of the total.

Table 6: Recorded training-step costs for the default mathematical run on eight H200 GPUs. Component timers can overlap or nest.

Recorded operation	Seconds per step	Total hours
Total training step	788.9	21.91
Rollout generation	211.1	5.86
Actor update	474.4	13.18
Old-policy log probabilities	48.6	1.35
Reference-policy log probabilities	46.4	1.29
Advantage computation	2.4	0.07
Rollout weight update	3.9	0.11

ORPG first obtains the full gradient of each policy objective. For two gradients in $\mathbb{R}^P$, reconciliation then uses the three independent Gram quantities $a = \|g_1\|^2$, $b = \|g_2\|^2$, and $d = g_1^\top g_2$. Computing these quantities and forming the final vector combination each take $O(P)$ arithmetic; the coefficient calculation is $O(1)$. With evenly distributed parameter shards over D devices, the local vector operations take $O(P/D)$. The compatible coefficients can be computed from the Gram quantities using

$$V = \|v_q\|, \qquad t_i = (1-\alpha) + \alpha \frac{S}{V} n_i^{q-1}, \qquad \omega_i = \frac{S t_i}{\sqrt{t_1^2 a + t_2^2 b + 2 t_1 t_2 d}}, \tag{18}$$

which gives $\mathcal{R}_+ = \omega_1 g_1 + \omega_2 g_2$. No parameter-space matrix or differentiation through the coefficients is needed. The distributed geometry reduction aggregates the three Gram quantities; full-gradient acquisition, FSDP synchronization, and the optimizer perform their own computation and communication. Storing the two objective gradients uses $O(P)$ memory. The recorded actor-update timer covers the policy update as a whole, including gradient acquisition and reconciliation.

C Evaluation protocols

C.1 Prompt construction and evaluation details

Dataset sizes and calibration. Table 7 lists the training, calibration, and evaluation splits. The three Alpaca subsets are disjoint. Mathematical evaluation contains 6,060 problems in total and generates four responses per problem.

Table 7: Dataset sizes. Mathematical training uses DeepScaleR preview prompts.

Dataset	Use	Count
Alpaca	Training	50,978
Alpaca	Reward calibration	512
Alpaca	Evaluation	512
HH-RLHF	Evaluation	8,520
PKU-SafeRLHF	Evaluation	8,211
AIME-24	Evaluation	30
AMC-22-23	Evaluation	83
MATH	Evaluation	5,000
Minerva-Math	Evaluation	272
OlympiadBench	Evaluation	675

Mathematical training prompts. Each DeepScaleR training example supplies one user message. The exact content construction is:

```
{problem}
Please reason step by step, and put your final answer within
\boxed{}.
```

The problem text is stripped of leading and trailing whitespace before appending the instruction. The reference answer is stored separately for reward evaluation and is not part of the user message.

Helpfulness–safety messages. The policy receives the dataset-provided user/assistant message sequence through its native tokenizer chat template, with `add_generation_prompt=True`. Evaluation preserves the final user request. When a prompt exceeds 512 tokens after template application, the adapter first removes the oldest complete conversation turns; if the remaining user message is still too long, it retains a token suffix that fits the prompt budget. This maintains a valid user-started, user-ended conversation.

Helpfulness–safety decoding. Evaluation uses one response per prompt, temperature 0.7, top-$p = 1.0$, and a maximum of 1024 generated tokens. Reward scoring uses a maximum sequence length of 2048. All three datasets are evaluated in full for each seed. Their means are computed separately and then averaged with equal dataset weights. The policy generation and the two reward-model evaluations use their respective tokenizer interfaces.

Repeated-run aggregation. For each metric, let m_k denote a complete run's result. We report $\bar{m} = N^{-1} \sum_k m_k$ and sample standard deviation $s = \sqrt{\sum_k (m_k - \bar{m})^2/(N-1)}$. A macro result is constructed within each run before computing its standard deviation. Results in both settings use three runs with seeds 42, 43, and 44.

Budget-level example. The initial policy's recorded AIME-24 evaluation illustrates the role of the shorter budgets in HV. At 2048, 4096, and 8192 tokens, accuracy is 18.33%, 33.33%, and 57.50%, with mean response lengths of 1971, 3571, and 5761 tokens. The three-point HV is 0.2955, compared with an area of 0.1706 for the 8192 point alone. The additional points measure answer quality available at lower realized costs.

C.2 Mathematical evaluation and hypervolume

For dataset D, each prompt has four sampled responses. At budget b, the accuracy and mean length are

$$a_{D,b} = \frac{1}{4|D|} \sum_{x \in D} \sum_{k=1}^{4} \mathbf{1}\{\text{response } (x,k) \text{ is correct at } b\}, \qquad \ell_{D,b} = \frac{1}{4|D|} \sum_{x \in D} \sum_{k=1}^{4} L_{x,k,b}.$$

All responses remain in the denominator, including unparseable answers. The shorter-budget views use exact token prefixes of the same generated responses. With $e_{D,b} = 1 - \min(1, \max(0, \ell_{D,b}/8192))$, define

$$\mathrm{HV}_D = \mathrm{Area}\left(\bigcup_{b \in \{2048,4096,8192\}} [0, e_{D,b}] \times [0, a_{D,b}] \right).$$

The main table reports $a_{D,8192}$ and HV_D. Avg is the equal-weight mean over the five datasets. Each repetition is summarized before calculating its mean and sample standard deviation; dataset standard deviations are not averaged to obtain a macro standard deviation.

Table 8 provides the budget-specific macro accuracy and length measurements used to compute HV. The primary accuracy comparison uses the 8192-token budget; HV summarizes the union area defined above.

D Component comparisons and parameter sensitivity

D.1 Definitions of objective-wise comparison rules

All rules below act on the separate reward-specific policy gradients g_1, g_2 before the shared regularization contribution. Sum uses $g_1 + g_2$ for every pair. PCGrad uses the symmetric conflict projection in Algorithm 1 when $g_1^\top g_2 < 0$ and the direct sum otherwise. Thus, PCGrad is the compatible-coordination-off variant in the HS setting. The conflict-resolution-off variant uses Equation (7) for

Table 8: Budget-specific measurements used to compute hypervolume. Accuracy (Acc) is in percent; length (Len) is in tokens.

Method	2048		4096		8192	
	Acc ↑	Len ↓	Acc ↑	Len ↓	Acc ↑	Len ↓
Base	$43.76_{\pm 0.34}$	$1434_{\pm 2}$	$53.66_{\pm 0.48}$	$2232_{\pm 1}$	$65.52_{\pm 0.70}$	$3165_{\pm 12}$
GRPO	$51.09_{\pm 0.66}$	$1299_{\pm 54}$	$\mathbf{60.40}_{\pm 0.48}$	$1556_{\pm 70}$	$60.69_{\pm 0.52}$	$1568_{\pm 66}$
GDPO	$51.94_{\pm 0.19}$	$\mathbf{1192}_{\pm 19}$	$59.00_{\pm 0.41}$	$\mathbf{1385}_{\pm 36}$	$59.35_{\pm 0.12}$	$\mathbf{1392}_{\pm 38}$
GD^{2}PO	$\mathbf{52.31}_{\pm 1.27}$	$1204_{\pm 19}$	$59.84_{\pm 1.03}$	$1439_{\pm 33}$	$60.26_{\pm 0.85}$	$1456_{\pm 29}$
ORPG	$47.04_{\pm 0.32}$	$1406_{\pm 3}$	$59.00_{\pm 0.27}$	$2076_{\pm 8}$	$\mathbf{66.66}_{\pm 0.26}$	$2554_{\pm 22}$

compatible gradients and the direct sum for conflicting gradients. Removing both components gives Sum. The CAGrad and Aligned-MTL rows use their respective joint-gradient constructions within the same objective-wise policy interface (Liu et al., 2021; Senushkin et al., 2023).

D.2 Mathematical component comparisons

The four versions share the mathematical reward definitions and objective-wise advantage construction, including length advantages centered within the correct-response subset. Without compatible coordination retains correctness-priority conflict projection and directly sums compatible gradients. Without conflict resolution retains compatible coordination and sums conflicting gradients. Without either component sums the two objective gradients in every case.

Table 9: Mathematical component comparisons. Accuracy and length use the 8192-token budget; HV aggregates the three budget points.

Update	Accuracy (%) ↑	Length ↓	HV ↑
ORPG	$\mathbf{66.66}_{\pm 0.26}$	$2554.14_{\pm 21.83}$	$0.5226_{\pm 0.0017}$
Without compatible coordination	$66.15_{\pm 0.19}$	$2447.87_{\pm 30.61}$	$\mathbf{0.5270}_{\pm 0.0025}$
Without conflict resolution	$65.46_{\pm 0.55}$	$2456.62_{\pm 18.64}$	$0.5203_{\pm 0.0026}$
Without either component	$66.04_{\pm 0.14}$	$\mathbf{2399.56}_{\pm 5.39}$	$0.5236_{\pm 0.0021}$

Table 9 shows that the complete rule achieves the highest full-budget accuracy. It exceeds the version without conflict resolution by 1.20 percentage points, connecting correctness-priority projection to improved answer quality when compatible coordination is retained. Its accuracy also exceeds the versions without compatible coordination and without either component by 0.51 and 0.62 percentage points. The full rule uses longer responses to attain this accuracy. The version without compatible coordination achieves the highest HV, while direct summation gives the shortest responses. These comparisons show how the components affect the accuracy–cost trade-off under the shared correctness-first objective construction.

D.3 Sensitivity to compatible coordination parameters

We vary the reference exponent q and maximum mixing strength λ individually around the shared default $(q, \lambda) = (0.5, 0.25)$. The q scan uses $\{0.25, 0.5, 0.75\}$ at $\lambda = 0.25$; the λ scan uses $\{0.125, 0.25, 0.5\}$ at $q = 0.5$. The scans share their default point, giving five configurations per setting. Scores follow the main evaluation protocol: full-set mean@1 on 17,243 HS prompts, and four samples per problem on 6,060 mathematical problems with exact-prefix budgets of 2048, 4096, and 8192 tokens. Means and sample standard deviations are computed across the three runs after dataset-level macro aggregation.

Tables 10 and 11 show that the main gains persist across the tested parameter range. Useful ranges from 5.482 to 5.589 and Harmless from 6.806 to 6.904, exceeding the strongest external baseline on each metric in Table 1. Mathematical full-budget accuracy ranges from 65.43% to 66.66% and HV from 0.5205 to 0.5243; all five configurations exceed the external training baselines in Tables 2 and 3 on both metrics.

Table 10: Sensitivity on helpfulness–safety. Scores are macro averages over the three evaluation sets.

q	λ	Useful ↑	Harmless ↑
0.25	0.25	$5.5197_{\pm 0.0005}$	$6.8409_{\pm 0.0012}$
0.5	0.25	$\mathbf{5.5891}_{\pm 0.0004}$	$\mathbf{6.9038}_{\pm 0.0014}$
0.75	0.25	$5.5124_{\pm 0.0004}$	$6.8248_{\pm 0.0011}$
0.5	0.125	$5.4816_{\pm 0.0017}$	$6.8064_{\pm 0.0015}$
0.5	0.5	$5.5542_{\pm 0.0012}$	$6.8620_{\pm 0.0009}$

Table 11: Sensitivity on mathematics: full-budget accuracy and length, and three-budget hypervolume.

q	λ	Accuracy (%) ↑	Length ↓	HV ↑
0.25	0.25	$65.58_{\pm 0.33}$	$2425.42_{\pm 38.69}$	$0.5227_{\pm 0.0023}$
0.5	0.25	$\mathbf{66.66}_{\pm 0.26}$	$2554.14_{\pm 21.83}$	$0.5226_{\pm 0.0017}$
0.75	0.25	$66.60_{\pm 0.51}$	$2613.33_{\pm 12.82}$	$\mathbf{0.5243}_{\pm 0.0043}$
0.5	0.125	$65.43_{\pm 0.29}$	$\mathbf{2413.83}_{\pm 14.08}$	$0.5205_{\pm 0.0028}$
0.5	0.5	$65.52_{\pm 0.49}$	$2434.32_{\pm 11.65}$	$0.5218_{\pm 0.0036}$

The default achieves the highest HS scores and full-budget mathematical accuracy among these configurations. Its exponent $q = 0.5$ changes the reference amplitude ratio to the square root of the original norm ratio, retaining magnitude information while moderating large-norm contributions. The shared $\lambda = 0.25$ provides a moderate coordination strength supported by both task settings. Figure 5 shows the local response to each parameter. In mathematics, $q = 0.75$ gives the highest three-budget HV, while the default gives the highest full-budget accuracy. Table 12 further shows that the accuracy ranking changes with the token budget: $q = 0.25$ leads at 2048 tokens, $\lambda = 0.125$ at 4096, and the default at 8192. These outcomes describe how the parameters adjust the accuracy–cost trade-off across budgets.

Table 12: Mathematical sensitivity at each token budget. Accuracy is in percent and length in tokens.

		2048		4096		8192	
q	λ	Acc. ↑	Length ↓	Acc. ↑	Length ↓	Acc. ↑	Length ↓
0.25	0.25	$\mathbf{47.77}_{\pm 0.40}$	$\mathbf{1374.0}_{\pm 1.7}$	$59.31_{\pm 0.44}$	$\mathbf{2015.2}_{\pm 11.0}$	$65.58_{\pm 0.33}$	$2425.4_{\pm 38.7}$
0.5	0.25	$47.04_{\pm 0.32}$	$1406.3_{\pm 3.2}$	$59.00_{\pm 0.27}$	$2075.9_{\pm 7.9}$	$\mathbf{66.66}_{\pm 0.26}$	$2554.1_{\pm 21.8}$
0.75	0.25	$47.26_{\pm 0.61}$	$1390.6_{\pm 3.2}$	$59.33_{\pm 0.36}$	$2079.0_{\pm 19.7}$	$66.60_{\pm 0.51}$	$2613.3_{\pm 12.8}$
0.5	0.125	$47.34_{\pm 0.49}$	$1408.5_{\pm 3.1}$	$\mathbf{60.55}_{\pm 0.77}$	$2035.7_{\pm 14.5}$	$65.43_{\pm 0.29}$	$\mathbf{2413.8}_{\pm 14.1}$
0.5	0.5	$47.27_{\pm 0.44}$	$1380.7_{\pm 2.0}$	$59.44_{\pm 0.21}$	$2031.8_{\pm 2.9}$	$65.52_{\pm 0.49}$	$2434.3_{\pm 11.6}$

E Training dynamics and measurements

E.1 Additional training-gradient analysis

The measurements cover all 100 optimizer steps for the four component versions. Figure 4 presents gradient cosine and Useful RMS in the main text. Table 13 summarizes these signals together with conflict and projection rates over the same three training intervals.

Figure 6(a) shows when conflicting gradient pairs occur. The two versions retaining compatible coordination encounter conflicts later in training, while the other two record zero conflicts. Figure 6(b) shows that ORPG projects those pairs, whereas Without conflict resolution leaves them unprojected. A positive step-average cosine can coexist with conflicts on individual updates within that step.

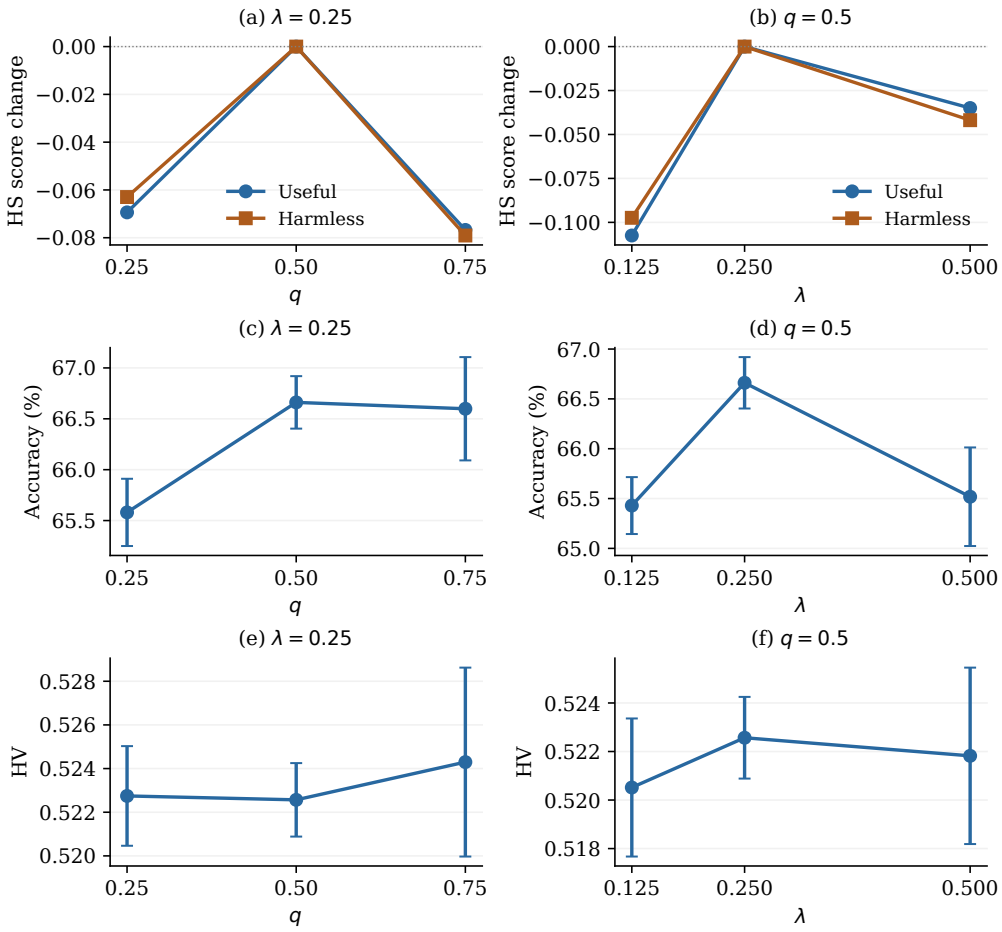

Figure 5: Parameter sensitivity with one parameter varied at a time. (a–b) HS score changes relative to the default; raw scores appear in Table 10. (c–d) Full-budget mathematical accuracy. (e–f) Three-budget mathematical HV.

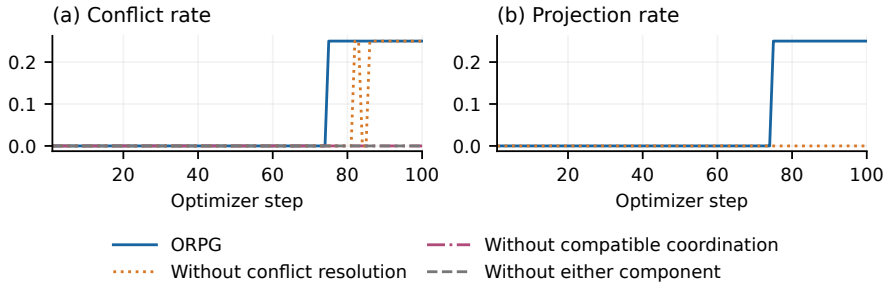

Figure 6: Unsmoothed conflict and projection statistics. (a) Conflict rates for all four component versions; the two zero-conflict curves overlap. (b) Projection rates for ORPG and Without conflict resolution.

Figure 7 supplements these relationships with the angle between the original sum and the compatible output. The statistic uses their normalized vector difference; conflict-branch calls contribute zero. The plot compares the two versions that apply compatible coordination.

Table 13: Training-stage gradient relationships and Useful advantage RMS. All versions use the same three step intervals.

Update	Steps	Cosine	Conflict	Projection	Useful RMS
ORPG	1–33	0.661	0.000	0.000	0.490
	34–66	0.584	0.000	0.000	0.500
	67–100	0.265	0.191	0.191	0.857
Without compatible coordination	1–33	0.690	0.000	0.000	0.492
	34–66	0.616	0.000	0.000	0.488
	67–100	0.522	0.000	0.000	0.507
Without conflict resolution	1–33	0.661	0.000	0.000	0.493
	34–66	0.576	0.000	0.000	0.484
	67–100	0.398	0.125	0.000	0.759
Without either component	1–33	0.692	0.000	0.000	0.493
	34–66	0.619	0.000	0.000	0.485
	67–100	0.523	0.000	0.000	0.504



Figure 7: Compatible-branch rotation for the two versions retaining compatible coordination. Values are unsmoothed step aggregates.

E.2 Training reward and advantage measurements

Training rewards are the calibrated Useful and Harmless values computed on the rollout batch. Evaluation reports raw reward-model scores on held-out prompts. The two quantities share objective meanings but have different scales and samples. For display, all reward curves average rewards across training runs at each step and then apply a trailing five-step mean. No variability band is inferred from temporal smoothing.

The component advantage RMS is the square root of the mean squared advantage over valid response tokens, recorded separately for each objective. All four component versions use shared-scale normalization and exact-zero handling. The values in Table 14 average steps 81–100 within each trajectory and then average actual training runs. The Useful RMS measurement is interpreted together with the observed Useful reward growth.

Table 14: Late-training calibrated rewards and Useful advantage RMS (steps 81–100).

Method	Useful	Harmless	Useful RMS
GDPO	1.1568	1.4582	–
GD^{2}PO	1.1577	1.4620	–
Without either component	1.1532	1.7368	0.5207
Without compatible coordination	1.1738	1.7676	0.5278
Without conflict resolution	1.4924	1.9990	0.8600
ORPG	**1.5217**	**2.0051**	0.8868