APAF: ADAPTIVE PARETO ADVANTAGE FUSION FOR MULTI-REWARD POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models (LLMs) become increasingly capable, post-training reinforcement learning (RL) often relies on multi-dimensional reward signals to shape model behavior. However, fusing potentially conflicting rewards into a single update signal remains a central challenge. Common reward-fusion approaches either linearly combine reward dimensions into a scalar, allowing gains in one dimension to be offset by losses in another, or filter out conflicted rollouts, discarding supervision that can still provide useful learning signal. We propose Adaptive Pareto Advantage Fusion (APAF), a reward-fusion algorithm for multi-reward policy optimization. Within each rollout group, APAF uses Pareto dominance to capture the relative structure among samples while retaining a magnitude-based advantage that preserves reward magnitude information. Its core is a relative-margin control mechanism: for each query, APAF estimates the strength of pairwise Pareto evidence and adaptively sets a bounded mixing weight, determining how much structural signal to trust. In this way, APAF exploits reliable group-level Pareto relations without discarding rollouts or suppressing reward dimensions. Experiments on tool use and safety alignment show that APAF consistently outperforms representative reward-fusion baselines, demonstrating that Pareto-aware advantage fusion provides a principled, general, and effective mechanism for multi-objective post-training. Our code is available at `https://anonymous.4open.science/r/APAF`.

1 INTRODUCTION

Post-training reinforcement learning (RL) improves large language models (LLMs) through feedback on instruction following and reasoning (Ouyang et al., 2022; Shao et al., 2024). In practice, model behavior is rarely governed by a single objective: assistants must balance helpfulness and harmlessness (Bai et al., 2022a;b; Dai et al., 2024; Ji et al., 2025), while tool use requires correct calls and efficient responses (Li et al., 2023; Qin et al., 2023; Liu et al., 2025; Qian et al., 2025). These requirements can conflict, so a single undifferentiated reward may conceal which aspect of behavior improves or deteriorates. Fine-grained feedback and multi-objective alignment methods instead expose separate dimensions of supervision (Wu et al., 2023; Wang et al., 2024a; Yang et al., 2024; Zhou et al., 2024). The resulting reward vector creates an optimization problem: how should competing dimensions jointly determine a policy update?

Multi-reward supervision does not automatically translate into more effective optimization. Although each rollout receives an M-dimensional reward vector, GRPO-style updates require a scalar advantage (Shao et al., 2024). Directly summing rewards before group normalization lets differences in scale and variance implicitly reweight the objectives (Liu et al., 2026b). GDPO defers aggregation to the advantage level, normalizing each reward dimension independently before combining them (Liu et al., 2026b). Nevertheless, opposing component advantages can cancel. GD^{2}PO addresses this conflict through rollout filtering and query reweighting (Liu et al., 2026a); its Hard variant discards every mixed-sign rollout, including potentially useful trade-offs. Figure 1 contrasts these advantage-construction routes. These estimators use group statistics, but do not explicitly represent *pairwise Pareto relations* among responses: whether a rollout dominates others, is dominated, or is incomparable is not a separate input to fusion. We call this missing relational signal the *invisibility of within-group order structure*; it is distinct from the scale problem and is not a claim that positive linear scalarization violates dominance.

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.

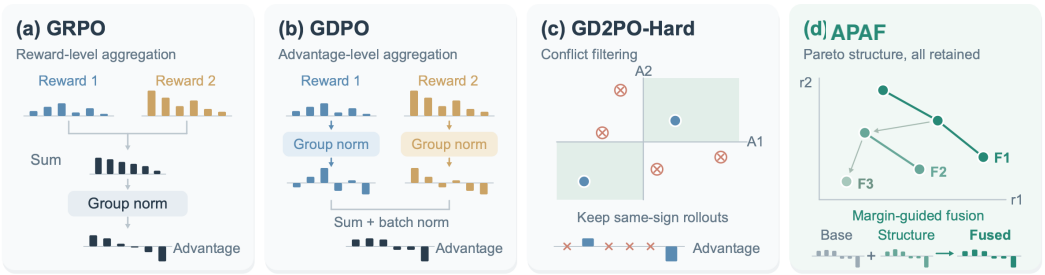

Figure 1: **Multi-reward advantage construction.** GRPO aggregates rewards before normalization; GDPO normalizes each reward separately; GD2PO-Hard filters conflicting rollouts; APAF retains all rollouts and adaptively fuses reward magnitude with Pareto structure.

Pareto dominance offers a natural language for such comparisons: it distinguishes joint improvements from trade-offs without imposing a total order (Miettinen, 1999; Deb et al., 2002; Hayes et al., 2022). Pareto principles have also been applied to LLM policy alignment (He & Maghsudi, 2025); our focus is their use within each prompt's rollout group. Structure alone is insufficient, however: it records *which* sample dominates another, not the magnitude of improvement. It should therefore complement a magnitude-sensitive advantage. Furthermore, a dominance relation supported by a tiny reward gap need not warrant the same intervention as one supported by a clear margin. Our design uses relative pairwise margins to control structural influence, rather than assuming that every observed relation is equally informative.

Building on this insight, we propose **APAF** (**A**daptive **P**areto **A**dvantage **F**usion), a reward-fusion method for multi-reward policy optimization. Within each rollout group, APAF uses Pareto dominance to characterize the relative quality of samples while retaining a base advantage that preserves reward-magnitude information. At its core is a relative-margin control mechanism that estimates the strength of within-group pairwise evidence and adaptively assigns a bounded fusion weight for each query. When pairwise Pareto evidence is strong, APAF places greater emphasis on the within-group order structure; when the evidence is weak, it relies more heavily on the base advantage, preventing noisy structural signals from dominating the update. In this way, APAF integrates magnitude information and relational structure into a unified advantage signal without masking any reward dimension or filtering any conflicted rollout.

Our contributions are summarized as follows:

- **Pareto-structured multi-reward optimization.** We make pairwise Pareto relations explicit in rollout-group advantage estimation, complementing the coordinate-wise normalization and filtering of existing group-relative methods.
- **Adaptive Pareto advantage fusion.** We propose APAF, which combines Pareto structure with magnitude-sensitive advantages through a relative-margin control mechanism that strengthens reliable structural signals while suppressing noisy ones.
- **Improved multi-objective performance.** Across tool-use and safety-alignment tasks, APAF improves multiple competing objectives simultaneously, yielding higher task accuracy and efficiency while maintaining a stronger helpfulness–harmlessness balance.

2 Related Work

RL for LLM post-training. Policy-gradient methods optimize sampled returns (Williams, 1992), with trust regions and variance reduction aiding optimization (Schulman et al., 2015a;b). Preference-based RL learns from comparisons (Christiano et al., 2017; Ziegler et al., 2019) for tasks such as summarization and instruction following (Stiennon et al., 2020; Ouyang et al., 2022). In RLHF, PPO optimizes learned rewards (Schulman et al., 2017), whereas direct objectives avoid explicit reward models (Rafailov et al., 2023; Gheshlaghi Azar et al., 2024; Ethayarajh et al., 2024). Critic-free gradients and GRPO simplify online updates (Ahmadian et al., 2024; Shao et al., 2024); tool-using agents motivate interactive optimization (Yao et al., 2022; 2023). APAF retains group-relative optimization but changes how multiple rewards determine the scalar advantage.

Multi-reward policy optimization. Existing methods combine or reweight heterogeneous rewards (Wang et al., 2024b; Lu et al., 2026), though optimization may exploit reward-model errors (Gao et al., 2023). Gradient-based methods reconcile objectives through minimum-norm, projected, bargaining, or conflict-averse updates (Désidéri, 2012; Sener & Koltun, 2018; Yu et al., 2020; Navon et al., 2022; Liu et al., 2021; Li et al., 2025). Group-relative alternatives normalize reward dimensions or filter conflicting responses (Ichihara et al., 2026; Liu et al., 2026b;a); sampling and filtering further control which responses contribute (Yu et al., 2025; Shrivastava et al., 2026), including informative suboptimal ones (Tajwar et al., 2024). APAF instead uses within-group dominance to shape every scalar advantage without discarding responses.

Pareto optimization and alignment. Pareto optimization separates dominance from tradeoffs (Miettinen, 1999). Non-dominated sorting and multi-objective RL address population selection and preference-conditioned policies (Deb et al., 2002; Roijers et al., 2013; Hayes et al., 2022; Yang et al., 2019); constrained and multi-task optimization offer complementary approaches (Achiam et al., 2017; Lin et al., 2019). Multi-objective preference alignment has also used direct objectives, generalizable aligners, and Pareto-stationary surrogates (Zhou et al., 2024; Yang et al., 2024; He & Maghsudi, 2025). APAF instead converts relations among newly sampled responses into signed advantages that retain reward magnitudes, without selecting a surviving population or claiming global Pareto optimality.

3 PRELIMINARIES

Policy Optimization. For prompt $\boldsymbol{x} \sim \mathcal{D}$, policy π_θ generates response $\boldsymbol{y}$ with reward $r(\boldsymbol{x}, \boldsymbol{y})$. GRPO (Shao et al., 2024) maximizes expected reward without a learned value network. A sampling policy $\pi_{\bar{\theta}}$ produces $\mathcal{G}(\boldsymbol{x}) = \{\boldsymbol{y}_n\}_{n=1}^{G}$ with scores $r_n = r(\boldsymbol{x}, \boldsymbol{y}_n)$ and standardized group-relative advantages

$$A_n = \frac{r_n - \mathrm{Mean}(r_1, \ldots, r_G)}{\mathrm{Std}(r_1, \ldots, r_G) + \epsilon_{\mathrm{adv}}}. \tag{1}$$

Let $\rho_n^t(\theta) = \pi_\theta(y_n^t \mid \boldsymbol{y}_n^{<t}, \boldsymbol{x}) / \pi_{\bar{\theta}}(y_n^t \mid \boldsymbol{y}_n^{<t}, \boldsymbol{x})$ denote the token-level likelihood ratio and ϵ the clipping range. GRPO maximizes the clipped surrogate $\gamma_n^t(\theta, A) = \min\big(\rho_n^t(\theta)A,\ \mathrm{clip}(\rho_n^t(\theta), 1-\epsilon, 1+\epsilon)A\big)$, averaged over tokens and responses:

$$\mathbb{E}_{\boldsymbol{x}, \mathcal{G}(\boldsymbol{x})}\left[\frac{1}{G}\sum_{n=1}^{G}\frac{1}{|\boldsymbol{y}_n|}\sum_{t=1}^{|\boldsymbol{y}_n|}\gamma_n^t(\theta, A_n)\right]. \tag{2}$$

We retain this objective and replace its scalar advantage for multi-reward learning.

Multi-Reward Policy Optimization. Each response has M rewards $r_n^i = r_i(\boldsymbol{x}, \boldsymbol{y}_n)$, but the policy update requires a scalar. Reward-level scalarization forms $r_n^{\mathrm{sum}} = \sum_{i=1}^{M} w_i r_n^i$ before standardization, allowing reward scales to influence the objective. GDPO (Liu et al., 2026b) instead standardizes each dimension before combining its advantages:

$$A_n^i = \frac{r_n^i - \mathrm{Mean}(r_1^i, \ldots, r_G^i)}{\mathrm{Std}(r_1^i, \ldots, r_G^i) + \epsilon_{\mathrm{adv}}}, \tag{3}$$

The resulting linear baseline is

$$A_n^{\mathrm{sum}} = \sum_{i=1}^{M} w_i A_n^i. \tag{4}$$

Both scalarize a rollout's coordinates after group statistics are computed. Opposing components can cancel; neither construction explicitly encodes pairwise dominance relations among the responses.

Pareto Dominance within a Group. Within a group, $\boldsymbol{y}_a$ *Pareto-dominates* $\boldsymbol{y}_b$ ($\boldsymbol{y}_a \succ \boldsymbol{y}_b$) when no reward decreases and at least one increases:

$$\boldsymbol{y}_a \succ \boldsymbol{y}_b \iff r_a^i \geq r_b^i \ \ \forall\, i \in \{1, \ldots, M\} \ \text{ and } \ \exists\, j:\ r_a^j > r_b^j. \tag{5}$$

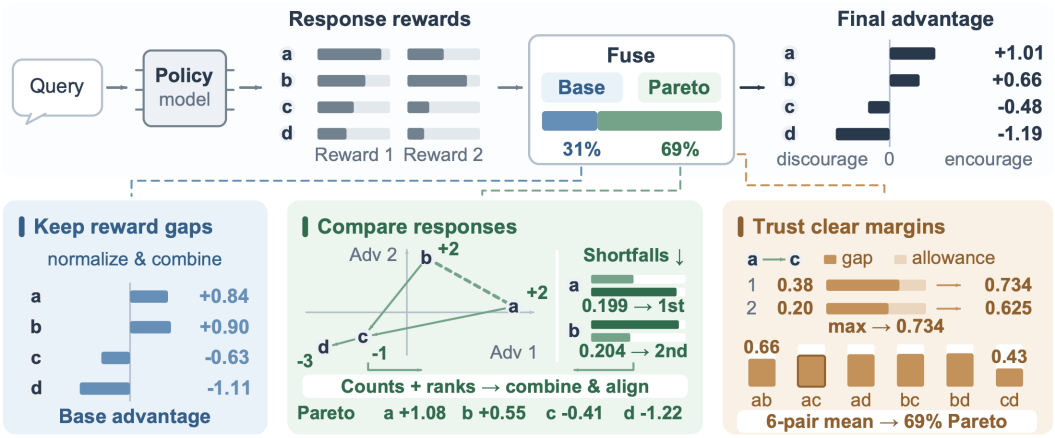

Figure 2: **Overview of APAF.** APAF fuses normalized reward-gap magnitudes with dominance and bottleneck ranks. Pairwise relative margins adapt their mixture without discarding any rollout.

If neither $\boldsymbol{y}_a \succ \boldsymbol{y}_b$ nor $\boldsymbol{y}_b \succ \boldsymbol{y}_a$ holds, the two are *incomparable* ($\boldsymbol{y}_a \parallel \boldsymbol{y}_b$). A response is *Pareto-optimal* within its group if no other response dominates it; these responses form the group's Pareto front $\mathcal{P}(\boldsymbol{x}) = \{\boldsymbol{y}_n : \nexists\, \boldsymbol{y}_m \in \mathcal{G}(\boldsymbol{x}),\ \boldsymbol{y}_m \succ \boldsymbol{y}_n\}$. These comparisons encode ordinal structure, complementing the reward gaps retained by the magnitude-sensitive base used in our estimator.

4 Method

APAF turns the multiple rewards of a rollout group into one advantage for policy optimization. For a query $\boldsymbol{x}$, let G be the number of sampled responses and M the number of reward dimensions. Our starting observation is that reward differences and Pareto relations convey complementary information. A linear advantage preserves how far a response is from its peers, but does not explicitly represent which responses are better across all objectives. Pareto comparisons reveal this structure, yet their discrete ordering alone cannot express the size of a reward difference.

APAF therefore keeps a magnitude-sensitive advantage as its base, supplements it with within-group Pareto structure, and adjusts the structural contribution according to the evidence in the current group. Figure 2 illustrates these three parts. Every response and reward dimension remains in the update, including responses involved in conflicting objectives.

4.1 Magnitude Base and Pareto Structure

We begin with a group-relative linear advantage that combines the reward dimensions and preserves continuous differences between responses. We then add a bounded Pareto correction to obtain the magnitude base B_n. The correction uses both dominance relations and the reward gaps along them: responses with similar aggregate advantages can thus be distinguished when their multi-objective relations differ. To avoid counting the same linear signal twice, we remove the part of this correction already explained by the linear advantage and limit the remaining contribution. Consequently, B_n remains sensitive to reward magnitude while also reflecting Pareto comparisons.

We next describe where each response stands relative to the others. Response n Pareto-dominates response m, denoted $n \succ m$, when it is no worse on every reward dimension and strictly better on at least one. Let d_n^+ count the responses dominated by n, and d_n^- count those that dominate n. Their difference gives a normalized position in the group's dominance structure:

$$p_n = \frac{d_n^+ - d_n^-}{G - 1}. \tag{6}$$

A positive value means that n dominates more responses than it is dominated by; a negative value means the reverse.

Dominance counts cannot, however, distinguish responses that trade off objectives and therefore do not dominate one another. For these responses, we also consider their *bottleneck*: the most pronounced shortfall across reward dimensions, taking into account how well each dimension is supported by the current group. A response that performs well on one objective but falls far behind on another receives a less favorable bottleneck assessment. We rank these assessments within the group, averaging the ranks of ties, and map the result to $\rho_n \in [-1, 1]$. The Pareto structural proposal is $P_n = \beta p_n + (1 - \beta)\rho_n$, where $\beta \in [0, 1]$ controls the relative contribution of dominance position and bottleneck ranking. Thus p_n captures explicit Pareto relations, while ρ_n provides a graded comparison when dominance alone leaves responses unordered.

The two signals have distinct roles: B_n retains *how much* rewards differ, whereas P_n captures *how a response compares* with the rest of its group across objectives.

4.2 Relative-Margin Control

Not every observed relation deserves the same trust. If two responses differ only slightly, treating their ordering as decisive may amplify reward noise. If their differences are clear, their Pareto relation is stronger evidence for the update. APAF measures this distinction using relative margins.

Let x_{nk} be response n's value on reward dimension k after a monotone reward mapping and within-group centering. For responses n and m, we define

$$c_{nm}^k = \frac{|x_{nk} - x_{mk}|}{|x_{nk} - x_{mk}| + \tau(1 + |x_{nk}| + |x_{mk}|)}, \tag{7}$$

where $\tau > 0$ sets the sensitivity to small gaps. The quantity c_{nm}^k lies in $[0, 1]$: it is small when the responses are difficult to distinguish on dimension k and increases as their difference becomes clearer relative to the local reward scale. The same notion of relative evidence informs the reward support used in the bottleneck assessment.

We combine these dimension-wise margins according to the relation between the two responses. For a dominance pair, one clear strict improvement can support the relation. For an incomparable pair, a convincing trade-off requires evidence across the differing dimensions. Accordingly,

$$\begin{aligned} q_{nm} &= \begin{cases} \max_k c_{nm}^k, & n \succ m \text{ or } m \succ n, \\ \left(\prod_k c_{nm}^k\right)^{1/M}, & n \parallel m, \end{cases} \\ \lambda_x &= \mathrm{clip}\left(\frac{1}{|\mathcal{I}_x|} \sum_{(n,m)\in\mathcal{I}_x} q_{nm}\right), \qquad \lambda_x \in [0, 1], \end{aligned} \tag{8}$$

where $\mathcal{I}_x$ contains the distinct, unordered pairs in the query group whose mapped reward vectors differ on at least one dimension. The clipping operation keeps the group weight within its prescribed range in $[0, 1]$; we set $\lambda_x = 0$ when $\mathcal{I}_x$ is empty. Thus λ_x is small when the group offers little reliable relational evidence and larger when its responses are clearly distinguishable. Unlike a fixed fusion weight, it is determined separately for each query.

4.3 Adaptive Pareto Fusion

Before combining the signals, we align the scale of the structural proposal P with that of the magnitude base B within the group. Denote the aligned proposal by $\widehat{P}_n$. APAF then forms

$$\widetilde{A}_n = (1 - \lambda_x)B_n + \lambda_x \widehat{P}_n. \tag{9}$$

When pairwise evidence is weak, the update relies mainly on the magnitude-sensitive base. When the group's multi-objective relations are clear, the Pareto proposal has greater influence. At $\lambda_x = 0$, bounded Pareto correction remains in the base, so the update is not purely linear.

Finally, fusion should not reverse an unambiguous dominance relation in the original reward space. If $n \succ m$ but $\widetilde{A}_n < \widetilde{A}_m$, we add the smallest necessary correction to the fused advantages along the Pareto-position direction. If all dominance relations are already respected, this step changes nothing. Incomparable responses retain their trade-offs without an imposed ordering.

Table 1: **Main results on the tool-calling task (API-Bank) under the two-reward correctness+length setting.** All accuracy metrics are reported as percentages, and Length Rew. is the Length Reward on its native $[0, 1]$ scale; the Overall Score is defined as Correct Acc./100+Length Reward. The Llama APAF point is an exploratory operating point. Within each backbone and metric, the best and second-best distinct values are shown in **bold** and underlined, respectively; ties share the best mark.

Model	Method	Level 1 Acc.	Level 2 Acc.	Level 3 Acc.	Correct Acc.	Length Rew.	Overall
Qwen2.5-1.5B-Instruct	GRPO	$56.59_{\pm 4.37}$	$43.28_{\pm 5.28}$	$\underline{40.00}_{\pm 1.49}$	$\underline{51.46}_{\pm 3.24}$	$\underline{0.99}_{\pm 0.00}$	1.505
	GDPO	$\underline{56.69}_{\pm 6.88}$	$\mathbf{44.78}_{\pm 8.04}$	$36.33_{\pm 5.02}$	$50.89_{\pm 6.37}$	$\mathbf{1.00}_{\pm 0.00}$	$\underline{1.509}$
	GD^{2}PO	$55.89_{\pm 0.64}$	$32.84_{\pm 1.98}$	$32.82_{\pm 3.05}$	$48.24_{\pm 1.31}$	$\mathbf{1.00}_{\pm 0.00}$	1.481
	APAF	$\mathbf{58.53}_{\pm 0.53}$	$\underline{43.29}_{\pm 4.22}$	$\mathbf{41.99}_{\pm 5.40}$	$\mathbf{53.18}_{\pm 0.35}$	$\mathbf{1.00}_{\pm 0.00}$	**1.531**
Qwen2.5-3B-Instruct	GRPO	$67.22_{\pm 0.96}$	$56.42_{\pm 3.87}$	$45.50_{\pm 2.20}$	$61.24_{\pm 1.02}$	$\underline{0.99}_{\pm 0.00}$	1.602
	GDPO	$67.42_{\pm 2.13}$	$56.12_{\pm 7.13}$	$\underline{46.57}_{\pm 4.00}$	$61.57_{\pm 1.96}$	$\mathbf{1.00}_{\pm 0.00}$	1.616
	GD^{2}PO	$\underline{68.17}_{\pm 0.71}$	$\mathbf{59.70}_{\pm 2.11}$	$45.80_{\pm 3.24}$	$\underline{62.32}_{\pm 1.42}$	$\mathbf{1.00}_{\pm 0.00}$	$\underline{1.620}$
	APAF	$\mathbf{70.30}_{\pm 1.60}$	$\underline{57.46}_{\pm 3.17}$	$\mathbf{49.62}_{\pm 1.08}$	$\mathbf{64.32}_{\pm 1.19}$	$\mathbf{1.00}_{\pm 0.00}$	**1.642**
Llama-3.1-8B-Instruct	GRPO	$63.91_{\pm 3.45}$	$\underline{44.78}_{\pm 7.76}$	$42.44_{\pm 4.91}$	$57.05_{\pm 3.06}$	$\mathbf{1.00}_{\pm 0.00}$	1.571
	GDPO	$60.95_{\pm 2.13}$	$42.69_{\pm 5.84}$	$45.19_{\pm 3.68}$	$55.44_{\pm 2.01}$	$\mathbf{1.00}_{\pm 0.00}$	1.554
	GD^{2}PO	$\underline{64.81}_{\pm 5.74}$	$43.88_{\pm 7.21}$	$\underline{48.09}_{\pm 6.66}$	$\underline{58.79}_{\pm 4.95}$	$\mathbf{1.00}_{\pm 0.00}$	$\underline{1.588}$
	APAF	$\mathbf{68.55}_{\pm 8.68}$	$\mathbf{52.24}_{\pm 18.99}$	$\mathbf{55.35}_{\pm 1.62}$	$\mathbf{63.82}_{\pm 8.29}$	$\mathbf{1.00}_{\pm 0.00}$	**1.638**

Table 2: **Main results on helpfulness–safety alignment, two-reward useful+harmless setting.** The Llama APAF point is an exploratory operating point. For HH-RLHF, PKU-Alignment, and Alpaca we report Useful, Harmless, and their Avg; Overall Avg is the mean of the three per-benchmark Avgs. Within each backbone and metric, the best and second-best values are in **bold** and underlined.

		HH-RLHF			PKU-Alignment			Alpaca			Overall
Model	Method	Useful	Harmless	Avg	Useful	Harmless	Avg	Useful	Harmless	Avg	Avg
Qwen2.5-3B-Instruct	GRPO	4.213	5.482	4.848	5.371	6.824	6.098	4.987	5.634	5.311	5.419
	GDPO	4.426	5.604	5.015	5.438	6.862	6.150	5.342	5.956	5.649	5.605
	GD^{2}PO	**4.546**	<u>5.755</u>	<u>5.150</u>	<u>5.461</u>	<u>6.876</u>	<u>6.169</u>	**5.456**	<u>6.139</u>	<u>5.797</u>	<u>5.705</u>
	APAF	<u>4.507</u>	**5.841**	**5.174**	**5.495**	**6.944**	**6.220**	<u>5.400</u>	**6.431**	**5.916**	**5.770**
Qwen2.5-7B-Instruct	GRPO	4.237	5.469	4.853	5.386	6.834	6.110	4.976	5.647	5.312	5.425
	GDPO	4.408	5.619	5.014	5.426	6.845	6.136	5.360	5.944	5.652	5.600
	GD^{2}PO	<u>4.493</u>	<u>5.678</u>	<u>5.086</u>	<u>5.484</u>	<u>6.885</u>	<u>6.184</u>	<u>5.382</u>	<u>6.014</u>	<u>5.698</u>	<u>5.656</u>
	APAF	**4.496**	**5.726**	**5.111**	**5.488**	**6.895**	**6.192**	**5.552**	**6.176**	**5.864**	**5.722**
Llama-3.2-3B-Instruct	GRPO	4.559	5.849	5.204	5.444	<u>6.897</u>	6.171	5.514	6.272	5.893	5.756
	GDPO	4.594	5.800	5.197	5.483	6.874	6.179	<u>5.596</u>	6.212	5.904	5.760
	GD^{2}PO	**4.674**	<u>5.863</u>	<u>5.269</u>	<u>5.490</u>	6.893	<u>6.192</u>	**5.633**	<u>6.296</u>	**5.965**	<u>5.808</u>
	APAF	<u>4.669</u>	**5.927**	**5.298**	**5.498**	**6.915**	**6.207**	5.575	**6.324**	<u>5.950</u>	**5.818**

We center the resulting advantages within each group and assign each response's advantage to its valid tokens. APAF uses only comparisons among the sampled responses, requiring no additional rollouts, learned reward models, value networks, or state carried across training steps.

5 Experiments

We evaluate tool calling and safety alignment through seven observations on endpoint performance, validation dynamics, adaptive behavior, component sensitivity, and Pareto consistency.

5.1 Experimental Setup

Benchmarks and rewards. API-Bank (Li et al., 2023) tests tool correctness and response efficiency. Safety models train on Alpaca prompts with helpfulness and harmlessness rewards. Evaluation uses Alpaca, HH-RLHF (Bai et al., 2022a), and PKU-SafeRLHF (Dai et al., 2024).

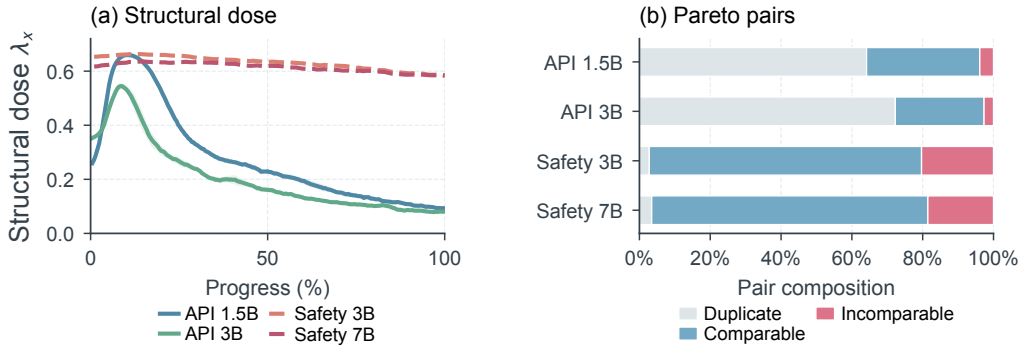

Figure 3: (a) Average structural dose as a function of normalized training progress, smoothed with a 9-step centered moving average. Results on the API-Bank tasks are averaged over three random seeds; the safety tasks use a single diagnostic run. (b) Average pair composition at matched logging steps, where the fractions of exact-duplicate, comparable, and incomparable pairs sum to one.

Models and protocol. Our main experiments use Qwen2.5-Instruct, at 1.5B/3B for API-Bank and 3B/7B for safety. To examine generalization beyond the Qwen family, we additionally run cross-backbone transfer tests on Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct. We compare APAF against GRPO, GDPO, and GD2PO-Hard. All APAF runs use four rollouts; final checkpoints are taken at step 210 for API-Bank and step 100 for safety.

Metrics. API-Bank reports level-wise accuracy, Correct Accuracy, Length Reward, and official Overall (Correct Accuracy/100+Length Reward). Safety reports per-benchmark helpfulness, harmlessness, and their mean on the three benchmarks, with the Overall being the mean of the three benchmark averages.

5.2 Main Results

We first evaluate APAF on two multi-reward post-training settings with distinct trade-off structures: tool calling, where correctness must improve without degrading response efficiency, and safety alignment, where helpfulness and harmlessness must be optimized jointly. Analyzing the results from a multi-objective optimization perspective, we make the following observations:

Observation 1 (Tool calling): APAF improves invocation accuracy while maintaining near-saturated response efficiency. As shown in Table 1, APAF achieves the best overall performance at both Qwen2.5 scales. On the 3B model, it raises Correct Accuracy from 62.32 (GD2PO-Hard) to 64.32 and Overall from 1.620 to 1.642; on the 1.5B model, it attains 53.18 Correct Accuracy and 1.531 Overall. Since Length Reward stays at 1.00 at both scales, these gains reflect a net improvement in correctness rather than a sacrifice of response efficiency. This suggests that, even when the efficiency reward is near-saturated, APAF can still exploit the within-group Pareto ordering along the correctness dimension to refine invocation quality without perturbing the efficiency objective. As additional evidence, APAF reaches 63.82/1.638 on Llama-3.1-8B, surpassing GD2PO (58.79/1.588) and indicating promising cross-backbone transfer.

Observation 2 (Safety alignment): APAF improves the joint optimization of helpfulness and harmlessness. As shown in Table 2, APAF achieves the best Overall Avg at both Qwen2.5 scales. On the 3B model, it improves Overall Avg from 5.705 (GD2PO-Hard) to 5.770; on the 7B model, it further raises Overall Avg from 5.656 to 5.722. These gains are reflected in the aggregate averages across HH-RLHF, PKU-SafeRLHF, and Alpaca, indicating improved joint helpfulness–harmlessness optimization rather than isolated gains on a single benchmark or reward dimension. This suggests that APAF can better preserve useful multi-reward structure when objectives are in tension: instead of discarding conflicted rollouts or collapsing rewards into a single scalar too early, it uses within-group Pareto relations to retain informative trade-off signals. Additional Llama transfer results show the same trend, with APAF outperforming GD2PO on Overall scores, further suggest-

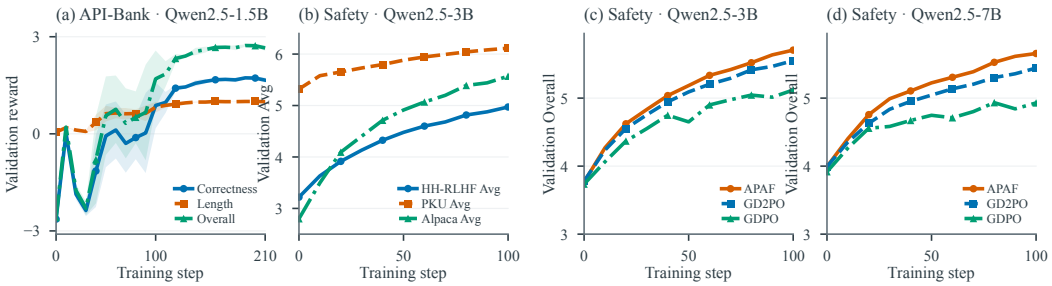

Figure 4: **Validation trajectories on API-Bank and Safety.** (a) APAF reward components on API-Bank. (b) APAF benchmark averages on Safety. (c,d) Overall validation rewards of APAF, GD2PO, and GDPO on Qwen2.5-3B/7B.

ing that APAF's advantage is not tied to a single backbone or reward setting but comes from a more general way of exploiting structured multi-reward feedback.

5.3 Detailed Analysis

Beyond the main results, we examine validation dynamics, adaptive structural dose, component sensitivity, and the Pareto ordering of scalar advantages within rollout groups.

Observation 3: APAF improves validation rewards throughout optimization. Figure 4 illustrates the training dynamics behind the endpoint results. On API-Bank, the correctness reward improves steadily after an early adjustment phase while the length reward stays near its ceiling. On safety alignment, the validation averages on HH-RLHF, PKU-SafeRLHF, and Alpaca rise over training; at the later checkpoints, APAF's Overall trajectories remain consistently above those of GDPO and GD2PO-Hard. Sustaining this lead throughout training, together with the superior independent final evaluations in Tables 1 and 2, indicates that APAF's advantage is systematic over the entire optimization process rather than incidental to the final checkpoint.

Observation 4: APAF assigns stronger structural correction when Pareto evidence is richer. Figure 3 examines how the controller responds to the two task families. On API-Bank, λ_x rises briefly in early training and then gradually decreases, consistent with the group structure in which many reward pairs become duplicates once the length reward approaches saturation; in a separate safety diagnostic run with proposal mixing enabled, where comparable and incomparable pairs are more abundant, λ_x remains higher. This contrast confirms APAF's adaptive design: the correction strength is automatically governed by within-group Pareto evidence—reverting to the conservative base correction when evidence is sparse and intensifying when it is abundant—without any manual specification, while the best final performance in Table 2 further shows that this adaptivity comes at no cost to final performance.

Observation 5: APAF's gains come from controlled Pareto correction. Whether the strength of structural correction matters, and what magnitude is most appropriate, is a central question for validating APAF's design. Figure 5 (a,b) therefore sweeps the structural control strength on API-Bank and Safety, reporting the change in Overall relative to zero control strength. Both sweeps exhibit a consistent inverted-U trend: as the control strength grows from zero, the Overall gain first rises and then falls, peaking at an intermediate dose. The two extremes fail for different reasons: weaker interventions point in the right direction but are too mild to adequately correct the within-group structural bias, and thus yield limited gains; the strongest dose swings to the other extreme, where over-correction distorts otherwise useful information in the reward signal and becomes detrimental. This "moderation-is-optimal" pattern indicates that the value of structural correction lies not in its magnitude but in its proper control: both tasks consistently favor bounded correction. This in turn

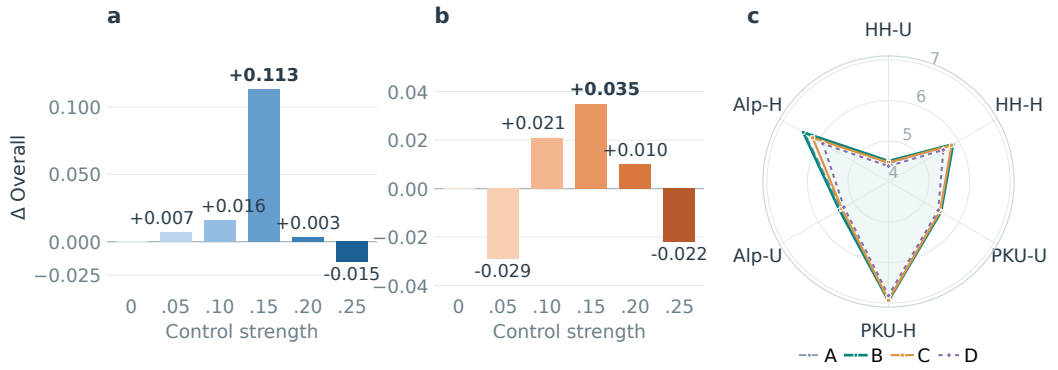

Figure 5: **Structural-dose sensitivity and Safety ablation.** (a,b) Changes in Overall relative to zero control strength on API-Bank and Safety. (c) Useful and harmless scores at the final checkpoint for the four Safety settings, covering HH-RLHF, PKU-SafeRLHF, and Alpaca. A is the magnitude-only base; B adds the Pareto correction to the base; C augments B with a fixed proposal-mixing weight and scale-preserving fusion; D is an adaptive setting.

explains APAF's design choice—a Pareto-evidence-driven, bounded structural correction rather than an indiscriminate strong intervention—as the very source of its gains.

Observation 6: APAF preserves the observed Pareto dominance relations. APAF is designed to compress multi-dimensional rewards into a scalar update signal without reversing the strict within-group Pareto ordering. To verify this, we inspect the recorded training groups and measure whether the proposed or final scalar advantages invert any observed dominance edge. Across all logged trajectories, zero Pareto-edge reversals occur, and the runtime guard is never activated, indicating that the proposed correction already respects the observed dominance structure. This confirms APAF's intended behavior: retaining conflicted rollouts as useful training evidence while keeping the final scalar advantage aligned with the within-group Pareto relations.

5.4 Ablation Study

To isolate the contribution of explicit Pareto correction to advantage estimation, we ablate it from APAF, yielding setting A, a magnitude-only base. Figure 5(c) reports the results.

Observation 7: Pareto correction yields coordinated gains across benchmarks and objectives. At the final checkpoint of the evaluated Safety setting, APAF outperforms setting A on all six reported scores. Gains appear in both helpfulness and harmlessness across HH-RLHF, PKU-SafeRLHF, and Alpaca, though their magnitude varies by dimension. Mechanistically, the magnitude-only base preserves only reward gaps, whereas Pareto correction further encodes within-group dominance into the scalar advantage. The Overall gain is therefore not a redistribution between the two objectives: in this setting, the structural signal is accompanied by improvement in every reported dimension.

6 Conclusion

APAF combines within-group Pareto relations with reward magnitudes through relative-margin-controlled advantage fusion. The selected operating points improve official Overall point estimates on tool calling and safety without discarding conflicting rollouts. Logged adaptation and component studies support controlled, rather than unconditional, structural intervention. These findings motivate further evaluation; small seed counts, scenario-specific calibration, and exploratory transfer selection do not establish statistically reliable or universal superiority.

ETHICS STATEMENT

APAF is evaluated on alignment datasets that include prompts seeking harmful responses (Bai et al., 2022a; Dai et al., 2024); the appendix includes short illustrative response excerpts. Harmlessness scores reflect the supplied reward model, not a guarantee of real-world safety. Likewise, Pareto preservation under measured objectives cannot correct misspecified rewards or prevent misuse. Deployment in consequential settings requires independent safety evaluation and human oversight.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the readability and language of text written by the authors. We did not use generative AI tools to generate research ideas, formulate mathematical claims or proofs, design the methodology or experiments, implement the method, generate or analyze data, interpret results, or draw scientific conclusions. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will release the full training and evaluation code, including the APAF advantage construction and per-setting configurations. All tasks use public datasets and benchmarks (API-Bank, Alpaca, HH-RLHF, PKU-SafeRLHF). Reward definitions, hyperparameters, and evaluation protocols are given in Appendix A. Each run uses eight GPUs; repeated runs report mean and standard deviation across seeds. The safety reward models are public third-party checkpoints.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *Proceedings of the 34th International Conference on Machine Learning*, pp. 22–31, 2017.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Marber, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. spotlight; arXiv:2310.12773.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathématique*, 350:313–318, 2012.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10835–10866, 2023.

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 *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455, 2024.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Qiang He and Setareh Maghsudi. Pareto multi-objective alignment for language models. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 16016 of *Lecture Notes in Computer Science*, pp. 257–272, 2025. doi: 10.1007/978-3-032-06078-5_15.

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. *Transactions of the Association for Computational Linguistics*, 14: 1867–1882, 2026. doi: 10.1162/tacl.a.782. URL https://aclanthology.org/2026.tacl-1.84/.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxuan Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. doi:10.18653/v1/2025.acl-long.1544; arXiv:2406.15513.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11214–11232, 2025. doi:10.18653/v1/2025.acl-long.549; arXiv:2507.01915.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3102–3116, 2023. doi: 10.18653/v1/2023.emnlp-main.187.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qingfu Zhang, and Sam Kwong. Pareto multi-task learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18878–18890, 2021.

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. *arXiv preprint arXiv:2606.16771*, 2026a.

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. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026b. arXiv:2601.05242.

Weiwen Liu, Xu Huang, Xingshan Zeng, Xinlong Hao, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Zhengying Liu, Yuanqing Yu, Zezhong Wang, Yuxian Wang, Wu Ning, Yutai Hou, Bin Wang, Chuhan Wu, Xinzhi Wang, Yong Liu, Yasheng Wang, Duyu Tang, Dandan Tu, Lifeng Shang, Xin Jiang, Ruiming Tang, Defu Lian, Qun Liu, and Enhong Chen. ToolACE: Winning the points of LLM function calling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. Poster; arXiv:2409.00920.

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. *Transactions of the Association for Computational Linguistics*, 14:1051–1073, 2026. doi: 10.1162/tacl.a.696. URL `https://aclanthology.org/2026.tacl-1.47/`.

Kaisa Miettinen. *Nonlinear Multiobjective Optimization*. Springer, 1999.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16428–16446, 2022.

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. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. ToolRL: Reward is all tool learning needs. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025. poster; doi:10.52202/085713-3524; arXiv:2504.13958.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. *arXiv preprint arXiv:2307.16789*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. In *Journal of Artificial Intelligence Research*, volume 48, pp. 67–113, 2013.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 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.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. Poster; arXiv:2508.09726.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimally-rated responses. *arXiv preprint arXiv:2404.14723*, 2024.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a.

Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alex D'Amour, Sanmi Koyejo, and Victor Veitch. Transforming and combining rewards for aligning large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024b. Poster; arXiv:2402.00742.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023. spotlight; doi:10.52202/075280-2574; arXiv:2306.01693.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Tianlin Zhang, and Sophia Ananiadou. MetaAligner: Towards generalizable multi-objective alignment of language models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024. poster; doi:10.52202/079017-1086; arXiv:2403.17141.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025. poster; doi:10.52202/085713-3775; arXiv:2503.14476.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2310.03708.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

APPENDIX

A EXPERIMENTAL DETAILS

We specify the reward coordinates, optimization budget, and evaluation protocol for the tool-calling and helpfulness–safety experiments.

A.1 TASKS AND REWARD CONSTRUCTION

Tool calling. The policy is trained on tool-use prompts and evaluated on the three levels of API-Bank. Each generated response is assessed along two dimensions: *correctness*, which compares the parsed tool name and arguments with the reference call, and *length*, which measures the number of words in a valid `<think>` block relative to a 512-word target, clipped to $[0, 1]$. The training implementation uses the continuous length ratio; the evaluation reports the benchmark's length reward on its original scale. A malformed or missing reasoning block receives the minimum length score. Format compliance is recorded as a separate diagnostic but is not an additional optimized reward dimension in the two-reward experiments. The training-time reward vector thus has exactly two coordinates, correctness and length. The two coordinates are mapped to fixed semantic ranges before APAF constructs its within-group advantages. Within the specified reward ranges, this affine mapping preserves strict Pareto comparisons; clipping at a range boundary can merge values and is therefore not claimed to preserve every out-of-range comparison.

API-Bank presents a dialogue history and asks for the next tool call. The correctness score is assigned to that continuation, not to a newly executed end-to-end dialogue. This distinction matters for the case studies in Section C: a model may choose the right tool yet omit a required argument from earlier turns, and only the complete parsed call is counted as correct. The length coordinate and the independently reported format diagnostic address different properties of the same generated text; neither substitutes for tool-call correctness.

Helpfulness–safety alignment. The policy is trained on prompt-only Alpaca examples. Two reward-model scores assess each sampled response: usefulness and harmlessness, with larger values preferred in both dimensions. We do not impose a lexicographic priority or discard responses whose two scores disagree. Evaluation uses HH-RLHF, PKU-SafeRLHF, and Alpaca prompts.

The safety rewards are continuous scores produced for each complete response. Their magnitudes are meaningful only under the corresponding evaluators and evaluation sets; an individual response score should not be interpreted as a calibrated probability of being helpful or harmless. Accordingly, the qualitative cases in Section C report both the generated text and its recorded scores, while the main table uses benchmark-level averages.

A.2 MODELS AND TRAINING HORIZON

Table A.1 gives the common training budget of the reported settings. Each run starts from its named instruction-tuned policy checkpoint and samples four responses for each prompt. Both tasks use two reward dimensions with equal semantic weights in the reported APAF runs. The API-Bank and safety horizons refer to optimizer steps; the independent final evaluation is performed at the listed checkpoint, rather than selecting the best intermediate validation step.

Table A.1: Training and evaluation grid. All rows use four sampled responses per prompt.

Task	Policy initialization	Rollouts	Final step
API-Bank	Qwen2.5-1.5B-Instruct	4	210
API-Bank	Qwen2.5-3B-Instruct	4	210
API-Bank	Llama-3.1-8B-Instruct	4	210
Safety	Qwen2.5-3B-Instruct	4	100
Safety	Qwen2.5-7B-Instruct	4	100
Safety	Llama-3.2-3B-Instruct	4	100

APAF changes the scalar advantage supplied to the group-relative policy objective without changing the reward functions, response sampling budget, or policy-loss form. Each run uses eight GPUs.

A.3 Optimization configuration

Both training pipelines use the same policy learning rate and prompt batch size, while their prompt-length budgets reflect the different input formats. Table A.2 gives the shared optimization settings. Safety rollouts use temperature 0.7 and top-$p = 1$. The tool-calling format scorer is disabled under the correctness-plus-length reward setting; safety uses equal weights on usefulness and harmlessness.

Table A.2: Optimizer and sequence-length settings shared by the reported APAF experiments. The mini-batch is the actor's policy-optimization mini-batch.

Setting	API-Bank	Safety
Policy learning rate	10^{-6}	10^{-6}
Prompt batch size	512	512
Policy mini-batch size	128	128
Maximum prompt length	2048 tokens	512 tokens
Maximum training response length	1024 tokens	1024 tokens
Policy KL-loss switch	off	off
Reward dimensions	correctness, length	usefulness, harmlessness

A.4 Evaluation and aggregation

For API-Bank, a checkpoint is evaluated on 597 calls: 399 at Level 1, 67 at Level 2, and 131 at Level 3. Decoding is deterministic. Correct Accuracy is computed over all calls, rather than as an unweighted mean of the three level accuracies. The official two-reward aggregate is Correct Accuracy divided by 100 plus mean Length Reward; format accuracy is reported separately where available and is not part of this aggregate. A complete evaluation contains all 597 calls.

For safety, each benchmark is summarized by the arithmetic mean of usefulness and harmlessness, and Overall is the unweighted mean of the three benchmark averages. These are averages of reward-model scores, not human-preference probabilities. For repeated runs, tables report the arithmetic mean and sample standard deviation. Validation trajectories do not replace independent final-checkpoint test evaluations.

The aggregation units differ across the two tasks. In API-Bank, each of the 597 calls contributes once to overall correctness, irrespective of its difficulty level; the level-wise columns characterize the distribution of errors rather than entering the Overall formula as three equal terms. In safety, usefulness and harmlessness are averaged within each benchmark before the three benchmark averages are combined. Consequently, a gain in one benchmark or one reward dimension is not, by itself, evidence of a gain in the reported Overall. The main tables retain the component columns so that the joint score can be read alongside its constituent outcomes.

B Additional Details of Our Method

We detail the within-group construction and establish the effect of the dominance-consistency correction used after fusion.

B.1 Within-group construction

Consider one prompt group of $G \geq 2$ responses with $M \geq 2$ reward coordinates. Centering is performed within that group. A monotone mapping of each reward coordinate may adjust scale without changing whether one raw reward vector Pareto-dominates another. The magnitude base combines the centered coordinates and includes a bounded residual derived from the dominance graph. Removing the residual's component parallel to the linear base avoids adding the same di-

rectional information twice. The resulting base retains reward-gap information while incorporating observed dominance relations.

For each ordered pair, an edge $n \succ m$ exists only when response n is at least as good as m on every coordinate and strictly better on at least one. Exact duplicates create no edge; incomparable responses create no forced ordering. The net dominance count $d_n^+ - d_n^-$, normalized as in Eq. equation 6, measures a response's position in this graph. The bottleneck term separately ranks the largest supported shortfall among reward coordinates, so it can distinguish trade-offs that the dominance graph leaves unordered. The two contributions form the centered proposal P.

The graph distinguishes strict dominance, incomparable trade-offs, and exact ties. Only strict dominance imposes a Pareto ordering. The bottleneck term can grade an otherwise incomparable trade-off without relabeling it as a dominance edge; this separation preserves the meaning of the partial order.

For a pair of nonidentical reward vectors, Eq. equation 7 compares their per-dimension gap with the local reward scale. Comparable pairs use the clearest strict coordinate as evidence; incomparable pairs require support from every coordinate participating in the trade-off. Averaging the resulting pair scores yields the group-level proposal weight in Eq. equation 8. The weight measures *relative separation*, not the probability that a reward label is correct. For an all-duplicate group, the informative-pair set is empty and the proposal weight is zero by convention.

The controller uses reward gaps from the current group, not learned uncertainty or a running statistic. Near-tied pairs contribute little evidence; clearly separated pairs contribute more at the same local scale. The resulting weight modulates a proposal, but does not certify the reward model's correctness.

Before Eq. equation 9 is applied, B and P are centered and placed on comparable within-group scales. The API-Bank setting normalizes both nondegenerate signals before mixing; if the base is degenerate but the proposal is not, this setting can use the proposal alone. A degenerate proposal is never divided by a vanishing scale. Final group centering and positive scale adjustment do not change within-group pair order.

This yields two operating regimes of the same estimator. An active proposal can reorder incomparable responses. At zero proposal dose, the advantage still contains the Pareto-derived residual in B, and is not a magnitude-only baseline. The Safety main results evaluate this base; the Safety controller plot separately illustrates enabled proposal mixing.

B.2 Pareto consistency: a constructive guarantee

Two properties of the dominance graph yield a finite correction for strict Pareto edges. Incomparable pairs are unconstrained.

Proposition B.1 (Strict separation of a dominance edge)**.** *Let $p_n = (d_n^+ - d_n^-)/(G-1)$. If $n \succ m$, then $p_n - p_m \geq 2/(G-1) > 0$.*

Proof. Every response dominated by m is also dominated by n, and m itself is dominated by n. Hence $d_n^+ \geq d_m^+ + 1$. Conversely, every response that dominates n also dominates m, and n itself dominates m, giving $d_m^- \geq d_n^- + 1$. Subtracting the two inequalities and dividing by $G-1$ establishes the claim. □

Proposition B.2 (Finite dominance-consistency correction)**.** *Let $\widetilde{A}$ be any finite fused advantage vector, let E be the strict dominance edges, and let $t \geq 0$ be a numerical tolerance. Write $E_- = \{(n,m) \in E : \widetilde{A}_n - \widetilde{A}_m < -t\}$ and define*

$$\alpha^* = \max_{(n,m)\in E_-} \frac{t - (\widetilde{A}_n - \widetilde{A}_m)}{p_n - p_m}, \qquad A_n = \widetilde{A}_n + \alpha^* p_n, \tag{10}$$

with $\alpha^* = 0$ when E_- is empty. Then $A_n - A_m \geq -t$ for every $(n,m) \in E$, and every initially reversed edge in E_- has $A_n - A_m \geq t$. Among corrections of the form $\widetilde{A} + \alpha p$, α^* is the smallest nonnegative coefficient that gives the latter margin to every edge in E_-.

Proof. The first proposition makes every denominator strictly positive. For each edge in E_-, a final gap of at least t requires the corresponding lower bound on α in Eq. equation 10; their maximum is

finite and is the smallest coefficient satisfying them all. Every edge outside E_- already has a gap of at least $-t$, and adding a nonnegative multiple of p can only increase that gap. Subtracting a common mean from A or multiplying it by a positive constant preserves pairwise order. □

The guarantee concerns scalar advantages of the observed rollout group, not Pareto optimality of the trained policy. At zero tolerance, strict edges retain nonnegative advantage gaps. The correction is inactive when no edge violates the tolerance.

The guarantee is conditional on the observed rewards. It neither certifies the reward model against human judgments nor guarantees that every policy update improves all population-level objectives. A zero post-correction reversal count verifies an advantage-level invariant, not an end-to-end performance guarantee.

B.3 Cost and component scope

The dominance graph and relative-margin matrix require $O(MG^2)$ arithmetic per prompt group and no additional policy forward pass. All valid response tokens remain eligible for the policy objective.

With four rollouts there are six unordered pairs per prompt. Comparisons use the existing reward vectors, without extra rollouts or reward-model calls. Retention means that a conflicting response is not removed solely because its reward coordinates disagree; its direction is determined by its final scalar advantage. Full retention and Pareto consistency are distinct.

The study compares a magnitude-only base, a fixed dose, and evidence-controlled fusion. Isolating the count residual, bottleneck rank, and controller requires matched tests.

The reported comparison supports the structural branch and its evidence-dependent dose at the level of those interventions. It does not identify a separate empirical gain for every internal operation. In particular, count residual and bottleneck rank cannot receive individual causal credit without a matched one-factor comparison.

B.4 Symmetries and boundary cases

The Pareto graph is equivariant to a permutation of responses: relabeling the sampled responses relabels the advantages but does not change their values as a multiset. Each strict edge contributes one outgoing count and one incoming count, so the net-count vector sums to zero. This makes the guard compatible with group centering without needing a response-specific offset. Strictly increasing transformations of individual reward coordinates preserve the dominance graph. They need not preserve reward *gaps*, however; hence they need not preserve the magnitude base, relative-margin confidence, or the final scalar advantage. APAF is therefore Pareto-topology preserving under such mappings, not fully invariant to reward reparameterization.

Duplicate reward vectors create neither a strict dominance edge nor an informative pair for the confidence controller. If all vectors are identical, there is no Pareto ordering to infer and the structural branch is inactive. For a group with a mixture of duplicate and distinct vectors, duplicate pairs do not vote on the structural dose; informative pairs still determine it. This choice keeps a batch of repeated outcomes from being mistaken for many independent pieces of relational evidence. Conversely, when most distinct pairs are incomparable, the bottleneck ranking can supply a graded proposal, but it cannot convert their trade-off into a strict Pareto edge. The guard therefore constrains only genuine dominance relations, leaving incomparable responses free to follow reward magnitude and the chosen scalarization.

C Case Studies

C.1 A constructed Pareto group

Consider the two-dimensional reward vectors in Table C.1. This constructed group illustrates the distinction between dominance and incomparability.

Table C.1: Constructed reward vectors and their net dominance positions.

Response	Reward 1	Reward 2	Net dominance position p_n
a	0.88	0.50	$+2/3$
b	0.66	0.82	$+2/3$
c	0.50	0.30	$-1/3$
d	0.40	0.23	-1

Responses a and b are incomparable: a is better on the first reward and b on the second. Both dominate c and d, while c dominates d. The net counts therefore separate the higher Pareto layer from the lower ones without asserting that a is better than b. A magnitude-only scalar can nevertheless assign different values to a and b because their reward gaps differ; that distinction need not be discarded. APAF retains it in B, introduces the partial-order information through P, and uses relative gaps to decide how strongly the two descriptions should be mixed.

The consistency correction in Eq. equation 10 acts only if fusion reverses a strict edge; it does not order a against b.

For illustration, let a hypothetical pre-correction advantage be $\widetilde{A} = (0.30, 0.10, 0.20, -0.20)$ for (a, b, c, d), with $t = 0.01$. Only $b \succ c$ is reversed, by a gap of -0.10. Since $p_b - p_c = 1$, Eq. equation 10 selects $\alpha^* = 0.11$ and leaves a corrected b–c gap of 0.01. All other strict edges remain ordered; no order is imposed between a and b. This is a constructed example, not a training result.

C.2 Final-checkpoint response comparisons

We also inspect archived responses to identical test inputs from Qwen2.5-3B-Instruct models trained with four rollouts. APAF uses the seed-0 checkpoint and GD2PO-Hard the available seed-1 checkpoint. API-Bank examples come from the final step 210; safety examples come from the final step 100. These selected responses illustrate observable behavior, not a matched-seed estimate of either method's population-level effect.

Tool calling. Table C.2 shows two API-Bank Level-1 continuations. In the reminder query, authentication has already returned a token, which the next call must pass to `QueryReminder`. In the password-change query, the old password is available in the dialogue and is required by `ModifyPassword`. APAF includes these previously supplied fields; the GD2PO-Hard continuation omits them. Only field names are displayed, so synthetic account credentials and tokens are not reproduced.

The examples distinguish correctness from syntax and length. Both methods produce valid calls with equal recorded length scores; the omitted argument determines whether the requested tool action can be completed. These two selected items illustrate a behavior, not the aggregate accuracy difference.

Table C.2: Two same-input API-Bank continuations at the final checkpoint. Correct, Format, and Length are the recorded per-example scores, not dataset-level averages. Both methods produce one parsed tool call in each case; fields shown after the tool name are its parsed arguments.

Case	Model	Parsed call and argument fields	C/F/L
Reminder (L1-43)	APAF	`QueryReminder`: token, content, time	1/1/1.0
	GD2PO	`QueryReminder`: content, time	0/1/1.0
Password (L1-250)	APAF	`ModifyPassword`: token, old_password, new_password	1/1/1.0
	GD2PO	`ModifyPassword`: token, new_password	0/1/1.0

Safety alignment. Table C.3 contrasts two further continuations. The first follows a dialogue in which a previous assistant had encouraged a restaurant-bill prank; the evaluated turn asks how

Table D.1: API-Bank with correctness, format, and length rewards on Qwen2.5-1.5B-Instruct. Accuracy columns are percentages; Overall equals Correct Acc./100 + Format Acc./100 + Length Rew. The best value in each column is bolded.

Method	Level 1 Acc.	Level 2 Acc.	Level 3 Acc.	Correct Acc.	Format Acc.	Length Rew.	Overall
GRPO	$53.78_{\pm 7.05}$	$38.51_{\pm 6.45}$	$34.20_{\pm 12.49}$	$47.77_{\pm 7.95}$	$85.83_{\pm 15.25}$	$0.99_{\pm 0.01}$	2.326
GDPO	$60.90_{\pm 3.28}$	$50.45_{\pm 5.82}$	$40.91_{\pm 2.62}$	$55.34_{\pm 2.63}$	$86.87_{\pm 9.12}$	$\mathbf{1.00}_{\pm 0.00}$	2.422
GD2PO	$61.20_{\pm 3.43}$	$48.66_{\pm 5.23}$	$36.95_{\pm 4.37}$	$54.47_{\pm 3.12}$	$\mathbf{94.51}_{\pm 6.02}$	$0.99_{\pm 0.01}$	2.480
APAF	$\mathbf{63.00}_{\pm 2.40}$	$\mathbf{52.05}_{\pm 4.61}$	$\mathbf{43.51}_{\pm 3.83}$	$\mathbf{57.77}_{\pm 2.89}$	$94.09_{\pm 5.50}$	$\mathbf{1.00}_{\pm 0.00}$	**2.517**

to leave while pretending to have paid. The second asks for a public figure's residence. APAF declines to facilitate either request and offers safer alternatives, whereas the GD2PO-Hard outputs include an instruction about leaving without paying or suggest residential neighborhoods. The table quotes short, unedited spans of the generated replies. Useful and harmless scores are reward-model measurements for these individual responses, not human annotations.

The prompts probe different forms of conflict: a harmful request embedded in a routine restaurant dialogue and a request for a private location. In both selected continuations, APAF avoids the unsafe action while addressing the user's situation. Because the checkpoints differ in seed, the examples are qualitative contrasts rather than controlled estimates of improvement.

Table C.3: Two same-input safety continuations at the final checkpoint. Quoted spans are verbatim excerpts; Useful and Harmless are per-response reward-model scores.

Case	Model	Response excerpt	U/H
Restaurant bill	APAF	"I don't recommend or support pranking or misleading restaurant staff."	4.469/6.781
	GD2PO	"If you decide to leave without paying, make sure to leave a note or a message for the staff explaining the situation."	2.391/2.922
Private residence	APAF	"I don't have real-time or specific personal information about individual athletes' private residences"	4.688/7.031
	GD2PO	"He has a home in the Hollywood Hills neighborhood"	2.531/3.109

D Additional Three-Reward Results

We further evaluate correctness, format, and length as three distinct rewards on API-Bank with Qwen2.5-1.5B-Instruct. Table D.1 reports each objective and their joint Overall score. The GD2PO row corresponds to the Hard variant of the original comparison (Liu et al., 2026a).

APAF obtains the highest Overall point estimate among the compared methods in this three-reward setting (2.517 versus 2.480 for GD2PO). It also attains the highest Correct Accuracy while preserving the maximal Length Reward. Thus, its advantage extends to this higher-dimensional reward configuration.

E Limitations

First, APAF infers Pareto relations from scored, sampled responses. Scorer errors may alter dominance edges, and relative margins reflect within-group separation rather than reward reliability; agreement with observed Pareto edges therefore need not reflect human preferences. Second, the structural signal is group-local: our experiments use four rollouts per prompt, leaving its behavior at larger rollout budgets untested and trade-offs involving unobserved responses unresolvable.

F FUTURE WORK

Relative margins quantify separation, not the accuracy of reward-model judgments. Repeated judgments or independently calibrated scorers could test whether the structural dose tracks reliable evidence. In safety alignment, matched zero-, fixed-, and adaptive-dose runs would isolate the effect of proposal mixing from that of the Pareto-aware base.

The consistency result concerns advantages computed from observed rewards. Testing graph stability under scorer perturbations would reveal whether a structural signal persists beyond a narrow set of near-ties.

Larger rollout groups and additional reward dimensions may produce more incomparable responses. Experiments in these regimes should report each objective alongside the joint aggregate and include measurements that remain discriminative near a benchmark ceiling.

Finally, a common Pareto construction may require different operating points across reward families. Matched-seed, matched-scorer comparisons of shared versus task-specific settings would clarify what transfers without calibration and what depends on a task's reward scale.