PREFERENCE-AWARE MULTI-OBJECTIVE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Alignment is an important component of post-training for modern foundation models and often involves multiple conflicting objectives with heterogeneous preferences specified by stakeholders. Failing to account for these conflicts and preferences can lead to undesirable trade-offs and unstable optimization. In practice, however, objective-level preferences may be only partially specified, e.g., through qualitative orderings among objectives rather than precise numerical weights. To address this challenge, we propose a preference-aware multi-objective alignment framework that models preferences over objectives through an admissible preference set, enabling a unified representation of both cardinal and ordinal preference information, including partial order constraints. Building on this framework, we develop PrEference-Consistent Cardinal-Ordinal Alignment (PECCO), an algorithm that mitigates gradient conflicts among objectives while preserving the specified preferences throughout optimization. We further establish convergence to a preference-aligned Pareto stationary point. Experiments on standard benchmarks show that PECCO achieves performance comparable to existing multi-objective preference-alignment methods while providing greater flexibility in accommodating diverse forms of objective-level preferences.

1 INTRODUCTION

Large language models (LLMs) are increasingly expected to align with multiple human-desired behaviors, such as helpfulness, safety, factuality, and conciseness (Bai et al., 2022a; Ouyang et al., 2022). Existing alignment methods, including RLHF and direct preference optimization, provide effective mechanisms for steering model behavior from human preference data (Christiano et al., 2017; Stiennon et al., 2020; Rafailov et al., 2023). However, practical alignment rarely involves a single criterion: different objectives may conflict, and improving one desired behavior can degrade another (Vamplew et al., 2018; Van Moffaert & Nowé, 2014; Wei et al., 2023). These challenges have motivated *multi-objective alignment*, which seeks to balance multiple competing alignment objectives. Recent approaches address this problem through model aggregation (Rame et al., 2023; Zhou et al., 2024), conditional policies (Wang et al., 2024a;b; Yang et al., 2024; Guo et al., 2024b), inference-time steering (Shi et al., 2024), and related mechanisms for navigating objective trade-offs.

Complementary to these approaches, a recent line of work formulates multi-objective alignment directly as multi-objective optimization (MOO), making the conflicts among objective gradients explicit during training. In particular, Chen et al. (2026a) adopts this formulation by modeling each alignment criterion as a separate preference loss, resulting in the vector-valued objective $F(x) = (f^1(x), \ldots, f^K(x))$ with solutions characterized by Pareto-type optimality conditions. This formulation naturally connects multi-objective alignment to a broad class of conflict-aware gradient methods developed for MOO and multi-task learning, including MGDA (Jean-Antoine, 2012), CAGrad (Liu et al., 2021), PCGrad (Yu et al., 2020), SMG (Liu & Nunes Vicente, 2019), MOCO (Fernando et al., 2022), and MODO (Chen et al., 2023), which seek update directions that balance competing objectives and mitigate gradient conflicts.

Despite this progress, existing methods remain limited in two related aspects: how preferences over objectives are modeled and how such preferences are preserved while resolving gradient conflicts. Most existing approaches assume that the desired trade-off can be specified through a precise numerical preference vector. Such cardinal information, however, may be difficult to elicit reliably in practice. Instead, it is often easier to specify qualitative ordinal information, e.g., some objectives have higher priority than others. This motivates a more flexible preference model that incorpo-

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.

rates uncertain or partially specified cardinal preferences together with reliable ordinal information. Beyond preference representation, the specified preference should also be respected consistently throughout optimization. This requirement is nontrivial because an update direction chosen to mitigate gradient conflicts may still deviate from the intended trade-off. These considerations motivate the central goal of our work:

> *To develop a multi-objective alignment framework that jointly accommodates cardinal and ordinal preference information, resolves conflicts among objective gradients, and preserves the prescribed preference structure throughout optimization.*

OUR CONTRIBUTIONS

- **Preference modeling.** Human preferences over alignment objectives are inherently multifaceted and contain diverse forms of information. Many existing approaches, such as Chen et al. (2026a), nevertheless collapse such preferences into a single prescribed numerical vector, potentially compressing away useful preference information. To model this richer preference information as possible, we introduce an *admissible preference set* that incorporates prescribed objective ordering, an optional numerical preference vector, and an explicit confidence level specifying how strongly the numerical preference should be enforced. We further characterize this admissible set as a convex hull of vectors, yielding a tractable feasible region for subsequent preference-aware optimization. Building on this formulation, we introduce the notion of a *preference Pareto stationary point*, which captures stationarity under the richer preference specification encoded by the admissible set. Importantly, the resulting preference model admits a clear geometric interpretation, which is supported by both our theoretical analysis and empirical observations.
- **Preference-consistent multi-objective alignment algorithm.** Building on the above preference model, we propose the *PrEference-Consistent Cardinal-Ordinal Alignment* (*PECCO*) algorithm, a *preference-consistent* and *conflict-avoiding* multi-objective alignment method. It steers optimization toward characterized objective preference by iteratively optimizing the multi-gradient combination weights over the admissible preference set, which makes the weights consistently reflect the specified preference at each step. Also, the update direction is obtained from the minimum-norm admissible multi-gradient combination, which yields a common descent direction and then avoids gradient conflicts among objectives. We further establish an $\mathcal{O}(\epsilon^{-2})$ gradient complexity for PECCO to reach an ϵ-preference Pareto stationary point.

2 RELATED WORK

Multi-Objective Optimization (MOO). Multi-task learning can be formulated as a MOO problem, where conflicting task gradients motivate balanced update directions (Sener & Koltun, 2018). MGDA (Jean-Antoine, 2012) provides a standard framework, with practical variants such as CAGrad (Liu et al., 2021) and PCGrad (Yu et al., 2020). Stochastic extensions include SMG (Liu & Nunes Vicente, 2019), MoDo (Chen et al., 2023), SDMGrad (Xiao et al., 2023), and MoCo (Fernando et al., 2022), while Zhang et al. (2024a) further establishes convergence under generalized smoothness. Preference-aware MOO further steers optimization toward desired trade-offs. Pareto MTL (Lin et al., 2019) uses preference vectors to target different Pareto regions, while Kaddani et al. (2017) models partial preference through feasible weight sets. PMGDA (Zhang et al., 2024b) imposes explicit preference constraints, and Chen et al. (2025a) optimizes a preference function.

LLM Alignment. Modern LLM alignment commonly learns from pairwise human preferences, either through reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022) or through direct preference optimization methods that avoid explicit reward-model training (Rafailov et al., 2023). Recent work has extended preference optimization to multiple objectives, including controllable alignment through preference-conditioned policies or objective weights (Wang et al., 2024a; Guo et al., 2024b; Wang et al., 2024b), as well as reward-free multi-objective preference optimization (Liu et al., 2025; Gupta et al., 2025; Chen et al., 2026a). These approaches establish the need for steerable trade-offs across competing alignment criteria, whereas our work focuses on incorporating structured preference information directly into the feasible gradient-combination set while retaining conflict-aware multi-objective optimization. We defer a more comprehensive discussion of related work to Appendix A.

3 PRELIMINARIES

3.1 REWARD-FREE ALIGNMENT AND DIRECT PREFERENCE OPTIMIZATION

Given a prompt $x \in \mathcal{X}$, an LLM parameterized by θ defines a policy $\pi_\theta(y \mid x)$ over responses $y \in \mathcal{Y}$. Due to the autoregressive nature of LLMs, the response probability can be factorized as $\pi_\theta(y \mid x) = \prod_{t=1}^{|y|} \pi_\theta(y_t \mid x, y_{<t})$. We denote by π_{ref} a fixed reference policy, typically taken as the pre-trained or supervised fine-tuned model. Classical RL-based preference alignment methods such as RLHF first learn a reward model from human preference data and then fine-tune the LLM through a KL-regularized RL objective, which can be computationally expensive and difficult to optimize. Direct Preference Optimization (DPO) (Rafailov et al., 2023) instead directly optimizes the policy from pairwise preference data, avoiding explicit reward modeling and RL training. DPO can be understood as a Bradley–Terry pairwise preference model in which the reward function is replaced by its closed-form representation induced by the optimal policy of the KL-regularized objective. Given a pairwise preference dataset $\mathcal{D} = \{(x, y_w, y_l)\}$, where y_w and y_l denote the preferred and dispreferred responses, respectively, the DPO objective is given by

$$f_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right), \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\beta > 0$ controls the strength of the deviation from the reference policy.

3.2 MULTI-OBJECTIVE ALIGNMENT

Equation 1 considers alignment with respect to a single preference objective. In practical LLM alignment, however, model responses are often evaluated along multiple criteria, such as helpfulness, harmlessness, factuality, and style. These criteria may induce distinct preference objectives $f^1(\theta), \ldots, f^K(\theta)$ that need to be optimized jointly. This naturally motivates formulating multi-objective LLM alignment as the following multi-objective optimization (MOO) problem.

$$\min_{\theta\in\mathbb{R}^d} F(\theta) := \big(f^1(\theta), \ldots, f^K(\theta)\big). \tag{2}$$

Since the objectives may conflict, a point simultaneously minimizing all objectives generally does not exist. Instead, the minimization above is understood in the *Pareto* sense, which characterizes non-dominated trade-offs among conflicting objectives. Related concepts are defined as follows.

Definition 3.1 (Pareto dominance)**.** For $\theta_1, \theta_2 \in \mathbb{R}^d$, we say that θ_1 *Pareto dominates* θ_2, if $f^k(\theta_1) \le f^k(\theta_2)$ for all $k \in [K]$ and $f^j(\theta_1) < f^j(\theta_2)$ for at least one $j \in [K]$. A point $\theta^\star \in \mathbb{R}^d$ is *Pareto optimal* if there exists no $\theta \in \mathbb{R}^d$ such that θ dominate θ^*. The set of all Pareto-optimal points is called the *Pareto set*, and its image under F is called the *Pareto front*.

In practice, however, the objective functions can be highly complex, making global Pareto optimality generally unattainable using only local oracle information such as gradients or Hessians. We therefore consider the tractable local notion of Pareto stationarity. We denote the probability simplex by $\Delta_K := \{w \in \mathbb{R}^K_+ : \mathbf{1}^\top w = 1\}$

Definition 3.2 (Pareto stationarity)**.** Suppose that $f^1, \ldots, f^K$ are differentiable. A point $\theta^\star \in \mathbb{R}^d$ is *Pareto stationary* if there exists $w \in \Delta_K$ such that $\nabla F(\theta^\star)w = \mathbf{0}$. It is called an ϵ-*Pareto stationary point* if $\min_{w\in\Delta} \|\nabla F(\theta)w\| \le \epsilon$.

However, these notions are limited for practical multi-objective alignment, where preferences may concern not only individual responses but also the relative importance of different objectives. By treating all Pareto stationary solutions as equally acceptable, the above definitions do not distinguish among solutions according to such objective-level preferences.

4 PREFERENCE-AWARE MULTI-OBJECTIVE ALIGNMENT

4.1 ADMISSIBLE PREFERENCE SET

Handling preference over objectives is a core challenge in multi-objective alignment. A common approach is to ask for a preference weight vector $\tilde{w} \in \Delta$ that quantifies the relative importance of

different objectives. However, such numerical preferences can be inherently uncertain. One often begins with a qualitative ordering of the objectives, while the exact weights consistent with that ordering may not be uniquely determined; for example, both $(w_1 = 0.7, w_2 = 0.3)$ and $(w_1 = 0.75, w_2 = 0.25)$ express a strong preference for objective f^1 over f^2. In other settings, one may provide only the ordering itself, without specifying any numerical weights. More importantly, once such an ordering is specified, it should be preserved consistently throughout the learning pipeline rather than treated as an informal prior.

Motivated by the above limitations, we introduce a preference set that jointly incorporates *ordinal* and *cardinal* preference information. Specifically, let $\mathcal{F} = \{f^1, \ldots, f^K\}$ denote the set of objectives, and suppose we specify a preference relation $\succeq$ over $\mathcal{F}$. This relation may define only a partial ordering, e.g., $f^1 \succeq f^2$ and $f^1 \succeq f^3$ indicate that objective f^1 is prioritized over both f^2 and f^3, while no preference between f^2 and f^3 is required. We define the corresponding ordinal preference set as

$$\mathcal{P}_{\mathrm{ord}} := \left\{ w \in \Delta : w_i \geq w_j \text{ whenever } f^i \succeq f^j \right\}. \tag{3}$$

When a numerical preference vector $\tilde{w} \in \Delta$ is additionally available, we integrate this cardinal information with the ordinal constraints through a confidence parameter $\lambda \in [0, 1]$, yielding the following admissible preference set.

Definition 4.1 (Admissible preference set)**.** Given objectives $\mathcal{F} = \{f^1, \ldots, f^K\}$, a preference relation $\succeq$, a numerical preference vector $\tilde{w} \in \Delta$ and a confidence parameter $\lambda \in [0, 1)$, the *admissible preference set* is defined as

$$\mathcal{P}_{\mathrm{adm}} := (1 - \lambda)\mathcal{P}_{\mathrm{ord}} + \lambda\tilde{w} = \{(1 - \lambda)w + \lambda\tilde{w} \mid w \in \mathcal{P}_{\mathrm{ord}}\}. \tag{4}$$

The parameter λ controls the confidence placed on the numerical preference vector $\tilde{w}$. When $\lambda = 0$, the formulation relies solely on the ordinal information and therefore does not require numerical preference weights. As λ increases, the admissible set concentrates around $\tilde{w}$ while preserving the specified ordering.

4.2 Extreme-Point Representation

Although intuitive, directly optimizing over $\mathcal{P}_{\mathrm{adm}}$ can be challenging because the ordinal constraints in $\mathcal{P}_{\mathrm{ord}}$ define a structured polytope within the high-dimensional probability simplex. Fortunately, the extreme-point representation of a polytope provides a convenient way to convert these ordinal constraints into a linear transformation of a standard simplex. Recall that a point $p \in \mathcal{P}_{\mathrm{ord}}$ is an *extreme point* if it cannot be expressed as a nontrivial convex combination of two distinct points in $\mathcal{P}_{\mathrm{ord}}$. Since $\mathcal{P}_{\mathrm{ord}}$ is a polytope, it is the convex hull of its finitely many extreme points.

Specifically, let $\{p^1, \ldots, p^M\}$ denote the extreme points of $\mathcal{P}_{\mathrm{ord}}$, and define $P_{\mathrm{ord}} := [p^1, \ldots, p^M] \in \mathbb{R}^{K \times M}$. Thus, by the Minkowski–Weyl theorem, the ordinal preference set can be rewritten as

$$\mathcal{P}_{\mathrm{ord}} = \mathrm{conv}\{p^1, \ldots, p^M\} = P_{\mathrm{ord}}\Delta,$$

where $\Delta = \{u \in \mathbb{R}_+^M : \mathbf{1}^\top u = 1\}$. Combining this representation with the cardinal preference $\tilde{w}$ yields the following representation of the admissible preference set.

Lemma 4.2 (Extreme-point representation)**.** *Define* $P_{\mathrm{adm}} := (1 - \lambda)P_{\mathrm{ord}} + \lambda\tilde{w}\mathbf{1}^\top \in \mathbb{R}^{K \times M}$*. Then the admissible preference set admits the equivalent representation*

$$\mathcal{P}_{\mathrm{adm}} = P_{\mathrm{adm}}\Delta = \{P_{\mathrm{adm}}u \mid u \in \Delta\}. \tag{5}$$

This representation replaces the explicit ordinal constraints with a standard simplex constraint on u. The matrix P_{adm} therefore serves as a linear map that encodes both the ordinal structure and the numerical preference, which will be particularly useful for developing efficient preference-aware optimization algorithms.

Table 1 presents several examples of ordinal preference sets together with their corresponding extreme points, while Figure 1 visualizes some representative feasible regions. For example, consider $\mathcal{P}_{\mathrm{ord}} = \{w \in \Delta : w_1 \geq w_2 \geq w_3\}$, whose extreme points are $(1, 0, 0)^\top$, $(\frac{1}{2}, \frac{1}{2}, 0)^\top$, and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^\top$. These extreme points progressively distribute probability mass over the top-ranked objectives: the

Table 1: Ordinal objective preferences and their extreme point representations.

Objective preference	$\mathcal{P}_{\text{ord}}$	$p^1, ..., p^M$
$f^1 \succeq f^2 \succeq f^3$	$\{w \in \Delta : w_1 \geq w_2 \geq w_3\}$	$(1,0,0)^\top,\ (\frac{1}{2},\frac{1}{2},0)^\top,\ (\frac{1}{3},\frac{1}{3},\frac{1}{3})^\top$
$f^1 \succeq f^2,\ f^1 \succeq f^3,$ $f^2 = f^3$	$\{w \in \Delta : w_1 \geq w_2,\ w_1 \geq w_3,$ $w_2 = w_3\}$	$(1,0,0)^\top,\ (\frac{1}{3},\frac{1}{3},\frac{1}{3})^\top$
$f^1 \succeq f^2, f^1 \succeq f^3$	$\{w \in \Delta : w_1 \geq w_2, w_1 \geq w_3\}$	$(1,0,0)^\top,\ (\frac{1}{2},\frac{1}{2},0)^\top,\ (\frac{1}{2},0,\frac{1}{2})^\top,\ (\frac{1}{3},\frac{1}{3},\frac{1}{3})^\top$
N/A	Δ	$(1,0,0)^\top,\ (0,1,0)^\top,\ (0,0,1)^\top$

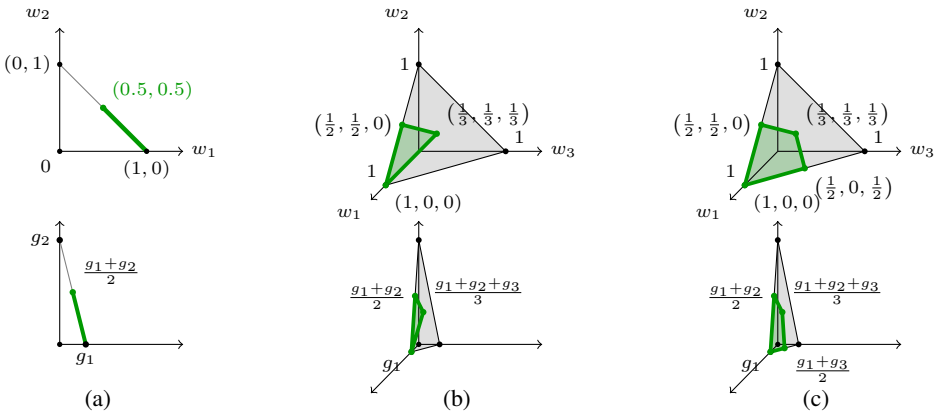


Figure 1: Ordinal preference sets in weight space (top) and their mappings to gradient space (bottom): (a) $w_1 \geq w_2$; (b) $w_1 \geq w_2 \geq w_3$; and (c) $w_1 \geq w_2$, $w_1 \geq w_3$, with $w_2 \parallel w_3$, denoting no ordinal relation is imposed between w_2 and w_3. See Appendix B for details and extension.

first concentrates entirely on f^1, the second assigns equal weight to f^1 and f^2, and the third assigns equal weight to all three objectives. Hence, every extreme point respects the prescribed ordering, and so does any convex combination of them. In this sense, the ordinal preference structure is encoded directly into the extreme-point representation. Using this representation, we next define a preference-aware notion of Pareto stationarity.

Definition 4.3 (ϵ-Preference Pareto stationary point)**.** Given objectives $f^1, \ldots, f^K$ and an admissible preference set $\mathcal{P}_{\text{adm}}$, a point θ is an *ϵ-preference Pareto stationary point* if $\min_{w \in \mathcal{P}_{\text{adm}}} \|\nabla F(\theta) w\| \leq \epsilon$, or equivalently, $\min_{u \in \Delta} \|\nabla F(\theta) P_{\text{adm}} u\| \leq \epsilon$.

5 Preference-Aware Multi-Gradient Descent

We introduce *PrEference-Consistent Cardinal-Ordinal alignment (PECCO)*, a preference-aware multi-objective alignment algorithm that mitigates conflicts among objective gradients while preserving the prescribed preference structure throughout optimization. PECCO builds on the classical Multiple Gradient Descent Algorithm (MGDA) (Zhang et al., 2024a), which constructs a common descent direction by balancing the gradients of multiple objectives. However, standard MGDA optimizes over the entire probability simplex and therefore does not explicitly account for prescribed preferences. Instead, we restrict the gradient-combination weights to the admissible preference set $\mathcal{P}_{\text{adm}}$. This leads to the following preference-constrained minimax problem:

$$\min_{w \in \mathcal{P}_{\text{adm}}} \max_{g} \langle \nabla F(\theta) w, g \rangle - \frac{1}{2}\|g\|^2. \tag{6}$$

The inner maximization determines the multi-gradient direction associated with a given admissible preference weight w, while the outer minimization selects the least-norm gradient combination among all weights consistent with the prescribed preference. Using the representation $\mathcal{P}_{\text{adm}} = P_{\text{adm}}\Delta$ from equation 5, the constraint $w \in \mathcal{P}_{\text{adm}}$ can be reparameterized as $w = P_{\text{adm}} u$ with $u \in \Delta$. Therefore, the preference-constrained subproblem becomes the standard simplex-

constrained problem $\min_{u\in\Delta}\max_g \langle \nabla F(\theta)P_{\mathrm{adm}}u, g\rangle - \frac{1}{2}\|g\|^2$. This allows us to leverage MGDA-type updates in the transformed preference space, and leads to the following PECCO iteration:

$$\begin{aligned}\text{(PECCO):}\quad u_{t+1} &= \Pi_\Delta\left(u_t - \gamma\left[(\nabla F(\theta_t)P_{\mathrm{adm}})^\top \nabla F(\theta_t)P_{\mathrm{adm}}u_t + \rho u_t\right]\right),\\ \theta_{t+1} &= \theta_t - \alpha_t \nabla F(\theta_t)P_{\mathrm{adm}}u_t.\end{aligned} \tag{7}$$

PECCO enjoys the following key technical properties for multi-objective alignment.

- **Conflict avoidance.** Conflicting objective gradients can cause an update that improves some objectives while degrading others. PECCO addresses this conflict within the set of trade-offs that are consistent with the prescribed preference. First, replacing the full simplex Δ by $\mathcal{P}_{\mathrm{adm}}$ restricts the feasible multi-gradient set from $\nabla F(\theta)\Delta$ to $\nabla F(\theta)\mathcal{P}_{\mathrm{adm}}$, thereby focusing the optimization on preference-admissible trade-offs, as illustrated in Figure 1. Second, the u_t-update approximately solves $\min_{u\in\Delta}\frac{1}{2}\left\|\nabla F(\theta_t)P_{\mathrm{adm}}u\right\|^2$, which finds the minimum-norm element g^* of the convex set $\nabla F(\theta_t)P_{\mathrm{adm}}\Delta$. By the projection optimality condition, the descent direction $d^* := -g^*$ satisfies $\langle d^*, g\rangle \le -\|g^*\|^2 < 0$ for every $g \in \nabla F(\theta_t)P_{\mathrm{adm}}\Delta$ whenever $g^* \neq 0$. Hence, d^* provides a common descent direction for all admissible multi-gradient combinations, resolving gradient conflicts while remaining consistent with the prescribed preference structure.
- **Preference consistency and accommodation of uncertainty.** PECCO preserves the prescribed preference structure throughout optimization while accommodating uncertainty in the numerical preference. For ordinal preferences, the multi-gradient is restricted to $\nabla F(\theta)\mathcal{P}_{\mathrm{ord}}$, ensuring that every admissible gradient combination respects the specified ordering among objectives. When a numerical preference vector $\tilde{w}$ is also available, the admissible set incorporates this cardinal information by biasing the multi-gradient toward the reference gradient $\nabla F(\theta)\tilde{w}$. The parameter $\lambda \in [0, 1)$ controls the confidence placed on this numerical preference. Together, the ordinal and cardinal information define the admissible multi-gradient set $\nabla F(\theta)\mathcal{P}_{\mathrm{adm}}$.

We establish the following convergence guarantee for PECCO under mild assumptions. The complete proof is provided in Appendix C.

Theorem 5.1 (Convergence of PECCO)**.** *Suppose that each objective f^k, $k \in [K]$, is L-smooth and lower bounded, i.e., $f^{k,\star} := \inf_\theta f^k(\theta) > -\infty$. Let the step sizes $\alpha, \gamma > 0$ be sufficiently small and independent of ϵ. Then, for any $\epsilon > 0$, by choosing $\rho = \mathcal{O}(\epsilon^2)$ and running PECCO for $T = \mathcal{O}(\epsilon^{-2})$ iterations, the generated iterates satisfy $w_t := P_{adm}u_t \in \mathcal{P}_{\mathrm{adm}}$, and $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\theta_t)w_t\|^2 \le \epsilon^2$.*

6 Experiments

We first introduce the datasets, baseline methods, and models used in the experiments presented in the main text. Additional datasets and implementation details are provided in Appendix D.

Task and Dataset. We evaluate PECCO on the Reddit Summary benchmark (Stiennon et al., 2020). For each Reddit post $\mathbf{x}$, the original dataset contains two candidate summaries $(\mathbf{y}^a, \mathbf{y}^b)$. These summaries are evaluated simultaneously along three aspects—*quality*, *conciseness*, and *faithfulness*—which serve as the three objectives in our experiments. The learning task therefore requires the simultaneous optimization of some or all of the three objectives. We reorganize the dataset into objective-specific preference pairs for multi-objective alignment as follows:

1. *Quality.* The quality preference is obtained using the GPT-2 summarization judge (Shi et al., 2024). The summary with the higher quality score is treated as y^+_{qual}, while the other is y^-_{qual}.
2. *Conciseness.* The conciseness preference is constructed according to summary length: the shorter summary is treated as y^+_{conc}, while the longer one is treated as y^-_{conc}.
3. *Faithfulness.* The faithfulness preference is obtained using the BART-based faithfulness judge (Chen et al., 2021), which scores each summary against the corresponding source post. The summary with the higher faithfulness score is treated as y^+_{faith}, while the other is y^-_{faith}.

Baselines and models. For methods that optimize objective weights, we consider DPO Loss Weight (DPO-LW) (Zhou et al., 2024), which directly combines objective gradients according

to a prescribed weight vector without explicitly resolving gradient conflicts, and RACO (Chen et al., 2026a), which primarily builds upon CAGrad (Liu et al., 2021). It searches for a conflict-averse common descent direction by solving the subproblem $\max_d \min_{k\in[K]} \langle \nabla f^k, d\rangle$ subject to $\|d - \sum_{k=1}^K \tilde{w}_k \nabla f^k\| \le c\|\sum_{k=1}^K \tilde{w}_k \nabla f^k\|$, where c is a constant and $(\tilde{w}_1, \ldots, \tilde{w}_K)$ denotes the prescribed numerical weight vector. We conduct experiments using Qwen3-4B-Instruct-2507 (Yang et al., 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024).

Metric. For each objective $k \in [K]$, we report the *margin* as the evaluation metric:

$$m_k = \mathbb{E}_{(\mathbf{x},\mathbf{y}_k^+,\mathbf{y}_k^-)\in\mathcal{B}_v}\left[\sigma\big(\log \pi_\theta(\mathbf{y}_k^+ \mid \mathbf{x}) - \log \pi_\theta(\mathbf{y}_k^- \mid \mathbf{x})\big)\right].$$

Following the definition in Chen et al. (2026a), a larger margin indicates that the trained policy assigns a stronger relative preference to the objective-specific preferred response.

6.1 Performance

We track the margin trajectories m_{qual} and m_{conc} for Qwen3-4B under prescribed numerical objective preferences $\tilde{w}_{\text{qual}} \in \{0.2, 0.5, 0.8\}$ and different confidence parameters, as shown in Figure 2 (complete results shown in Figure 14, 15). Pareto frontiers corresponding to additional weight pairs are provided in Figure 3a, illustrating the final objective trade-offs achieved by different methods. Unlike DPO-LW, which improves the model primarily along the more highly preferred objective at the expense of the other objective, the MOO-based methods PECCO and RACO exhibit more balanced improvements across objectives, consistent with the trade-off principle underlying multi-objective optimization. As shown in Figure 3b, we repeat the experiment using the Llama3-8B model, and observe similar results. We conduct additional experiments on the *quality-faithfulness* task, using Gemma models, and on the BeaverTails benchmark, respectively, and observe similar results, as shown in Appendix D.1.

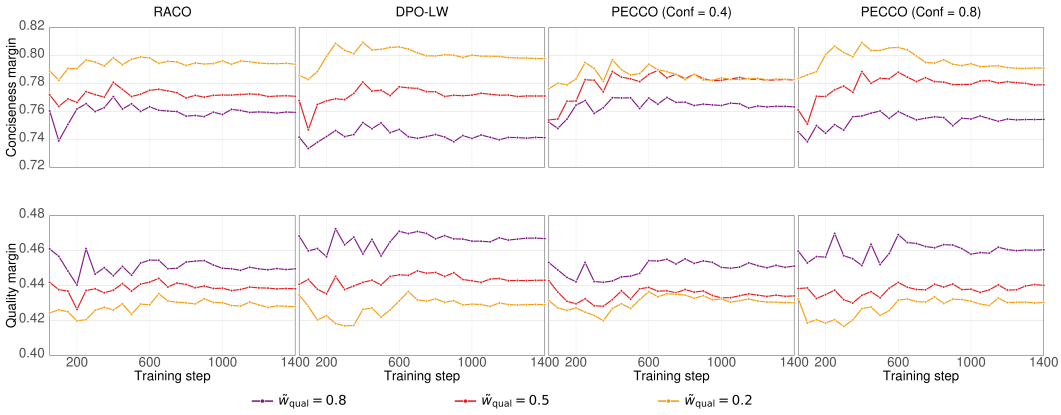
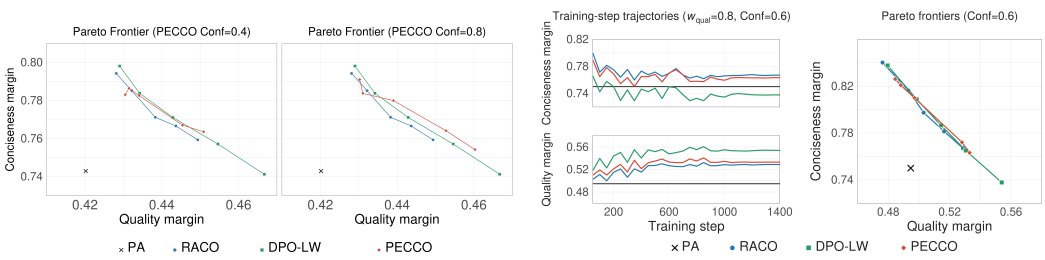


Figure 2: Training Trajectories of Qwen3-4B on the *Quality–Conciseness* Task



(a) Pareto Frontiers of Qwen3-4B

(b) Trajectories and Pareto Frontiers of Llama3-8B

Figure 3: Pareto Frontiers on Qwen3-4B and Performance Results on Llama3-8B

We further evaluate PECCO on a three-objective *quality–conciseness–faithfulness* alignment task on the Reddit Summary benchmark using the Qwen3-4B model. As shown in Table 2 (see Table 7 for the complete results), PECCO exhibits a clear pattern in the final margins that is highly consistent with the prescribed cardinal and ordinal preferences.

Preference	$(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}, \tilde{w}_{\text{faith}})$	Quality Margin	Conciseness Margin	Faithfulness Margin
$f^1 \succeq f^2 \succeq f^3$	$(0.6, 0.3, 0.1)$	0.4362	0.7548	0.5043
$f^1 \succeq f^2 \succeq f^3$	$(0.8, 0.15, 0.05)$	0.4410	0.7481	0.5017
$f^1 \succeq f^3 \succeq f^2$	$(0.8, 0.05, 0.15)$	0.4444	0.7428	0.5079
$f^2 \succeq f^1 \succeq f^3$	$(0.3, 0.5, 0.2)$	0.4266	0.7639	0.5106
$f^2 \succeq f^3 \succeq f^1$	$(0.2, 0.5, 0.3)$	0.4230	0.7664	0.5133
$f^3 \succeq f^2 \succeq f^1$	$(0.2, 0.3, 0.5)$	0.4258	0.7589	0.5257

Table 2: PECCO under representative three-objective linear preferences with confidence parameter $\lambda = 0.8$. Here, f^1, f^2, and f^3 denote quality, conciseness, and faithfulness, respectively.

6.2 Consistency and Conflict Avoidance

Preference consistency. To illustrate the effect of our preference modeling on the weights, we record the weight vector at each training step. As shown in Figure 4, PECCO provides a strict and geometrically interpretable control over the combining weights, with the sampled weights remaining concentrated within the prescribed admissible region. In contrast, RACO exhibits pronounced trajectories extending toward the original simplex vertices, whereas such extreme-point-directed patterns are substantially suppressed under PECCO. This observation is consistent with PECCO being less susceptible to distortions induced by large differences in objective-gradient magnitudes.

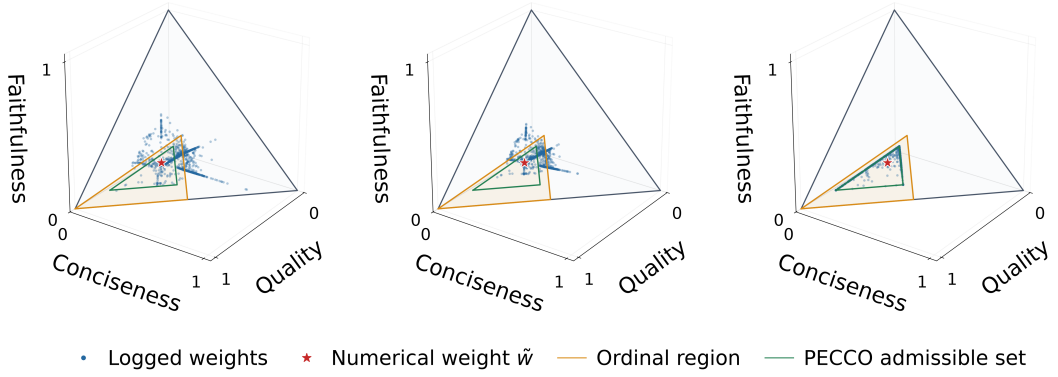


Figure 4: Training weight distributions of RACO (left: unnormalized; middle: normalized) and PECCO with confidence 0.4 (right). The preference is $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}, \tilde{w}_{\text{faith}}) = (0.5, 0.3, 0.2)$ with the corresponding ordering. Gray, orange, and green denote the simplex, ordinal region, and PECCO admissible set, respectively.

Conflict avoidance. Our method demonstrates a strong capability for avoiding gradient conflicts. As shown in Table 3, for the baseline methods, we measure the fraction of training updates in which the combined gradient conflicts with at least one original objective gradient. For PECCO, conflicts are instead evaluated with respect to the columns of $H = FP_{\text{adm}} = (h^1, \ldots, h^M)$, since these are the effective objective gradients directly involved in each optimization update. PECCO exhibits nearly zero conflict rates across all configurations, consistent with its conflict-avoidance mechanism.

$(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}})$	RACO	DPO-LW	PECCO (Conf $= 0.4$)	PECCO (Conf $= 0.8$)
$(0.20,\ 0.80)$	0.77%	10.99%	0.00%	0.00%
$(0.35,\ 0.65)$	0.07%	1.83%	0.00%	0.00%
$(0.50,\ 0.50)$	0.07%	0.85%	0.14%	0.00%
$(0.65,\ 0.35)$	0.28%	3.87%	0.00%	0.00%
$(0.80,\ 0.20)$	1.55%	9.65%	0.00%	0.00%

Table 3: Gradient conflict rates over 1,420 updates per configuration. A conflict occurs when the final combined direction has cosine similarity below -10^{-6} with at least one reference gradient. Reference gradients are the original objective gradients for RACO and DPO-LW, and the columns of $H = FP_{\text{adm}}$ for PECCO. Lower is better.

6.3 Ablation Studies

We conduct ablation studies on the confidence parameter λ in this section, while ablations on the iteration parameters γ and ρ are provided in Appendix D.2. As shown in Table 4 (see Table 8 for the complete results) and Figure 5, varying confidence parameter λ clearly leads to changes in the resulting objective trade-offs.

$(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.45, 0.55)$			$(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.80, 0.20)$		
Conf	Quality Margin	Conciseness Margin	Conf	Quality Margin	Conciseness Margin
0.0	0.4318	0.7842	0.0	0.4425	**0.7692**
0.2	0.4318	**0.7851**	0.2	0.4448	0.7641
0.4	0.4335	0.7842	0.4	0.4509	0.7617
0.6	0.4336	0.7803	0.6	0.4533	0.7571
0.8	**0.4368**	0.7815	0.8	**0.4590**	0.7504

Table 4: Ablation on confidence parameter λ under two cardinal preferences (Qwen3-4B).

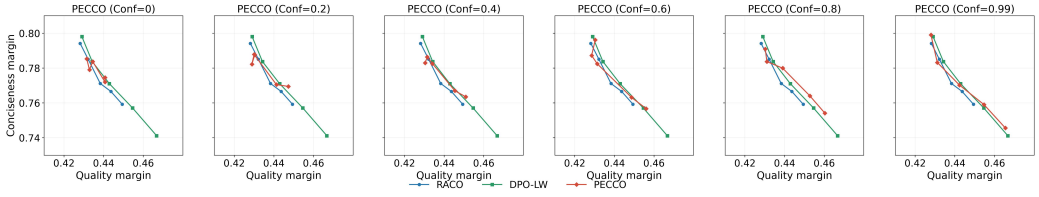


Figure 5: Qwen3-4B Pareto frontiers with different confidence parameters under $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ and corresponding ordinal preference with different confidence level.

7 Conclusion

In this work, we studied preference-aware multi-objective alignment. We introduced an admissible preference set that jointly models objective preferences from ordinal, cardinal, and confidence aspects. Building on this formulation, we proposed PECCO, which steers multi-objective alignment toward the specified preference. PECCO preserves preference consistency throughout optimization while avoiding gradient conflicts among objectives. We further established convergence to an ϵ-preference Pareto stationary point with $\mathcal{O}(\epsilon^{-2})$ gradient complexity. Our work also suggests several directions for future research in multi-objective alignment. While the current formulation represents preferences through polytopes, other geometric structures may capture user preferences in richer or more precise forms. Exploring such representations is an interesting direction for future work.

AI USE STATEMENT

We used generative AI tools only to polish the manuscript, improve readability, and assist with literature retrieval. We did not use generative AI tools to generate theoretical results or proofs, datasets, experimental designs, or interpretations of experimental results. All AI-assisted edits and literature-related suggestions were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *ArXiv Preprint: 2303.08774*, 2023.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. In *NeurIPS*, 2025.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *ArXiv Preprint: 1606.06565*, 2016.

M. Ansary and G. Panda. A modified quasi-Newton method for vector optimization problem. *Optimization. A Journal of Mathematical Programming and Operations Research*, 64(11):2289–2306, 2015.

A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, et al. A general language assistant as a laboratory for alignment. *ArXiv Preprint: 2112.00861*, 2021.

M. G. Azar, Z. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, pp. 4447–4455, 2024.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv Preprint: 2204.05862*, 2022a.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *ArXiv Preprint: 2212.08073*, 2022b.

C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996. doi: 10.1145/235815.235821.

L. Barrett and S. Narayanan. Learning all optimal policies with multiple criteria. In *ICML*, pp. 41–47, 2008.

H. Bonnel, A. N. Iusem, and B. F. Svaiter. Proximal methods in vector optimization. *SIAM Journal on Optimization*, 15(4):953–970, 2005.

David Bremner, Komei Fukuda, and Ambros Marzetta. Primal–dual methods for vertex and facet enumeration. *Discrete & Computational Geometry*, 20(3):333–357, 1998. doi: 10.1007/PL00009389.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *NeurIPS*, pp. 1877–1901, 2020.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *ArXiv Preprint: 2303.12712*, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, September 2023.

Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance, 2023.

Lisha Chen, Quan Xiao, Ellen Hidemi Fukuda, Xinyi Chen, Kun Yuan, and Tianyi Chen. Efficient First-Order Optimization on the Pareto Set for Multi-Objective Learning under Preference Guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 9443–9486. PMLR, October 2025a.

Peter Chen, Xi Chen, Wotao Yin, and Tianyi Lin. ComPO: Preference Alignment via Comparison Oracles, October 2025b.

Peter Chen, Xiaopeng Li, Ziniu Li, Xi Chen, and Tianyi Lin. Spectral policy optimization: Coloring your incorrect reasoning in GRPO. In *2nd AI for Math Workshop @ ICML 2025*, 2025c.

Peter Chen, Xiaopeng Li, Xi Chen, and Tianyi Lin. Reward-free Alignment for Conflicting Objectives, February 2026a.

Peter Chen, Xiaopeng Li, Ziniu Li, Wotao Yin, Xi Chen, and Tianyi Lin. Exploration vs exploitation: Rethinking rlvr through clipping, entropy, and spurious reward. In *ICLR*, 2026b.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5935–5941, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.475.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, pp. 4302–4310, 2017.

J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *ICLR*, 2024.

I. Das and J. E. Dennis. Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM Journal on Optimization*, 8(3): 631–657, 1998.

H. Dong, W. Xiong, B. Pang, H. Wang, H. Zhao, Y. Zhou, N. Jiang, D. Sahoo, C. Xiong, and T. Zhang. RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*, 2024.

L. M. G. Drummond and A. N. Iusem. A projected gradient method for vector optimization problems. *Computational Optimization and Applications*, 28(1):5–29, 2004.

L. M. G. Drummond and B. F. Svaiter. A steepest descent method for vector optimization. *Journal of Computational and Applied Mathematics*, 175(2):395–414, 2005.

M. Ehrgott. *Multicriteria Optimization*. Springer, 2005.

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Model alignment as prospect theoretic optimization. In *ICML*, pp. 12634–12651, 2024.

Heshan Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent stochastic approach, 2022.

J. Fliege and B. F. Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494, 2000.

J. Fliege, L. M. G. Drummond, and B. F. Svaiter. Newton's method for multiobjective optimization. *SIAM Journal on Optimization*, 20(2):602–626, 2009.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *ICML*, pp. 10835–10866, 2023.

J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. In *ICML*, pp. 1243–1252, 2017.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The Llama 3 herd of models. *ArXiv Preprint: 2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025a.

Jiacheng Guo, Ling Yang, Peter Chen, Qixin Xiao, Yinjie Wang, Xinzhe Juan, Jiahao Qiu, Ke Shen, and Mengdi Wang. GenEnv: Difficulty-aligned co-evolution between LLM agents and environment simulators. *ArXiv Preprint: 2512.19682*, 2025b.

S. Guo, B. Zhang, T. Liu, T. Liu, M. Khalman, F. Llinares, A. Rame, T. Mesnard, Y. Zhao, B. Piot, et al. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*, 2024a.

Y. Guo, G. Cui, L. Yuan, N. Ding, Z. Sun, B. Sun, H. Chen, R. Xie, J. Zhou, Y. Lin, et al. Controllable preference optimization: Toward controllable multi-objective alignment. In *EMNLP*, pp. 1437–1454, 2024b.

R. Gupta, R. Sullivan, Y. Li, S. Phatale, and A. Rastogi. Robust multi-objective preference alignment with online DPO. In *AAAI*, pp. 27321–27329, 2025.

Y. Y. Haimes, L. S. Lasdon, and D. A. Wismer. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1(3):296–297, 1971.

C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1), 2022.

D. Hernández-Lobato, J. Hernandez-Lobato, A. Shah, and R. Adams. Predictive entropy search for multi-objective Bayesian optimization. In *ICML*, pp. 1492–1501. PMLR, 2016.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging, October 2023.

Désidéri Jean-Antoine. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5):313–318, 2012. ISSN 1631-073X. doi: 10.1016/j.crma.2012.03.014.

J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, B. Chen, R. Sun, Y. Wang, and Y. Yang. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *NeurIPS*, pp. 24678–24704, 2023.

Sami Kaddani, Daniel Vanderpooten, Jean-Michel Vanpeperstraete, and Hassene Aissi. Weighted sum model with partial preference information: Application to multi-objective optimization. *European Journal of Operational Research*, 260(2):665–679, July 2017. ISSN 0377-2217. doi: 10.1016/j.ejor.2017.01.003.

A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *ArXiv Preprint: 2503.19786*, 2025.

C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos. Pareto-path multitask multiple kernel learning. *IEEE Transactions on Neural Networks and Learning Systems*, 26(1):51–61, 2014.

Z. Li, T. Xu, Y. Zhang, Z. Lin, Y. Yu, R. Sun, and Z-Q. Luo. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *ICML*, pp. 29128–29163, 2024.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qingfu Zhang, and Sam Kwong. Pareto Multi-Task Learning, December 2019.

B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu. Conflict-averse gradient descent for multi-task learning. In *NeurIPS*, pp. 18878–18890, 2021.

Q. Liu, J. Ruan, H. Li, H. Zhao, D. Wang, J. Chen, G. Wan, X. Cai, Z. Zheng, and T. Xu. AMoPO: Adaptive multi-objective preference optimization without reward models and reference models. In *ACL 2025 (Findings)*, pp. 8832–8866, 2025.

Suyun Liu and Luis Nunes Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning, 2019.

T. Liu, Y. Zhao, R. Joshi, M. Khalman, M. Saleh, P. J. Liu, and J. Liu. Statistical rejection sampling improves preference optimization. In *ICLR*, 2024a.

Z. Liu, M. Lu, S. Zhang, B. Liu, H. Guo, Y. Yang, J. Blanchet, and Z. Wang. Provably mitigating overoptimization in RLHF: Your SFT Loss is implicitly an adversarial regularizer. In *NeurIPS*, pp. 138663–138697, 2024b.

N. Mahdavi-Amiri and F. Salehi Sadaghiani. A superlinearly convergent nonmonotone quasi-Newton method for unconstrained multiobjective optimization. *Optimization Methods and Software*, 35(6):1223–1247, 2020.

R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, 2004.

Y. Meng, M. Xia, and D. Chen. SimPO: Simple preference optimization with a reference-free reward. In *NeurIPS*, pp. 124198–124235, 2024.

S. Merity, N. S. Keskar, and R. Socher. Regularizing and optimizing LSTM language models. In *ICLR*, 2018.

A. Messac and C. A. Mattson. Normal constraint method with guarantee of even representation of complete Pareto frontier. *AIAA Journal*, 42(10):2101–2111, 2004.

K. Miettinen. *Nonlinear Multiobjective Optimization*, volume 12. Springer Science & Business Media, 1999.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, pp. 27730–27744, 2022.

R. Y. Pang, W. Yuan, H. He, K. Cho, S. Sukhbaatar, and J. Weston. Iterative reasoning preference optimization. In *NeurIPS*, pp. 116617–116637, 2024.

S. Parisi, M. Pirotta, N. Smacchia, L. Bascetta, and M. Restelli. Policy gradient approaches for multi-objective sequential decision making. In *IJCNN*, pp. 2323–2330. IEEE, 2014.

Chanwoo Park, Ziyang Chen, Asuman Ozdaglar, and Kaiqing Zhang. Post-training llms as better decision-making agents: A regret-minimization approach. *ArXiv Preprint: 2511.04393*, 2025.

R. Park, R. Rafailov, S. Ermon, and C. Finn. Disentangling length from quality in direct preference optimization. In *ACL*, pp. 4998–5017, 2024.

S. Peitz and M. Dellnitz. Gradient-based multiobjective optimization with uncertainties. In *Numerical and Evolutionary Optimization Workshop*, pp. 159–182. Springer, 2017.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, pp. 2227–2237, 2018.

M. Pirotta and M. Restelli. Inverse reinforcement learning through policy gradient minimization. In *AAAI*, 2016.

F. Poirion, Q. Mercier, and J-A. Désidéri. Descent algorithm for nonsmooth stochastic multiobjective optimization. *Computational Optimization and Applications*, 68(2):317–331, 2017.

Ž. Povalej. Quasi-Newton's method for multiobjective optimization. *Journal of Computational and Applied Mathematics*, 255:765–777, 2014.

L. F. Prudente and D. R. Souza. A quasi-Newton method with Wolfe line searches for multiobjective optimization. *Journal of Optimization Theory and Applications*, 194(3):1107–1140, 2022.

S. Qu, M. Goh, and F. T. S. Chan. Quasi-Newton methods for solving multiobjective optimization. *Operations Research Letters*, 39(5):397–399, 2011.

S. Qu, C. Liu, M. Goh, Y. Li, and Y. Ji. Nonsmooth multiobjective programming with quasi-Newton methods. *European Journal of Operational Research*, 235(3):503–510, 2014.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, pp. 53728–53741, 2023.

R. Rafailov, J. Hejna, R. Park, and C. Finn. From r to $Qˆ*$: Your language model is secretly a Q-function. In *COLM*, 2024.

A. Rame, G. Couairon, M. Shukor, C. Dancette, J-B. Gaya, L. Soulier, and M. Cord. Rewarded soups: Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *NeurIPS*, pp. 71095–71134, 2023.

Yinuo Ren, Tesi Xiao, Michael Shavlovsky, Lexing Ying, and Holakou Rahmanian. COS-DPO: Conditioned One-Shot Multi-Objective Fine-Tuning Framework, June 2025.

D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

S. Schäffler, R. Schultz, and K. Weinzierl. Stochastic method for the solution of unconstrained vector optimization problems. *Journal of Optimization Theory and Applications*, 114(1):209–222, 2002.

O. Sener and V. Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, pp. 525–536, 2018.

A. Shah and Z. Ghahramani. Pareto frontier learning with expensive correlated objectives. In *ICML*, pp. 1919–1927. PMLR, 2016.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *ArXiv Preprint: 2402.03300*, 2024.

R. Shi, Y. Chen, Y. Hu, A. Liu, H. Hajishirzi, N. A. Smith, and S. S. Du. Decoding-time language model alignment with multiple objectives. In *NeurIPS*, pp. 48875–48920, 2024.

Ruizhe Shi, Runlong Zhou, and Simon Du. The crucial role of samplers in online direct preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 39255–39285, 2025.

Y. Song, G. Swamy, A. Singh, J. Bagnell, and W. Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *NeurIPS*, pp. 12243–12270, 2024.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. In *NeurIPS*, pp. 3008–3021, 2020.

F. Tajwar, A. Singh, A. Sharma, R. Rafailov, J. Schneider, T. Xie, S. Ermon, C. Finn, and A. Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *ICML*, pp. 47441–47474, 2024.

Y. Tang, Z. Guo, Z. Zheng, D. Calandriello, R. Munos, M. Rowland, P. H. Richemond, M. Valko, B. Pires, and B. Piot. Generalized preference optimization: A unified approach to offline alignment. In *ICML*, pp. 47725–47742, 2024.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv Preprint: 2302.13971*, 2023.

P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20(1):27–40, 2018.

K. Van Moffaert and A. Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pp. 6000–6010, 2017.

L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec. TRL: Transformers reinforcement learning, 2020.

B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, pp. 8228–8238, 2024.

H. Wang, Y. Lin, W. Xiong, R. Yang, S. Diao, S. Qiu, H. Zhao, and T. Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *ACL*, pp. 8642–8655, 2024a.

J. Wang, Y. Hu, C. K. Wai Yu, C. Li, and X. Yang. Extended Newton methods for multiobjective optimization: Majorizing function technique and convergence analysis. *SIAM Journal on Optimization*, 29(3):2388–2421, 2019.

K. Wang, R. Kidambi, R. Sullivan, A. Agarwal, C. Dann, A. Michi, M. Gelmi, Y. Li, R. Gupta, A. Dubey, et al. Conditional language policy: A general framework for steerable multi-objective finetuning. In *EMNLP 2024 (Findings)*, pp. 2153–2186, 2024b.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In *NeurIPS*, 2025.

A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How does LLM safety training fail? In *NeurIPS*, pp. 80079–80110, 2023.

Mingkang Wu, Devin White, Evelyn Rose, Vernon Lawhern, Nicholas R. Waytowich, and Yongcan Cao. Multi-Task Reward Learning from Human Ratings, June 2025.

Peiyao Xiao, Hao Ban, and Kaiyi Ji. Direction-oriented multi-objective learning: Simple and provable stochastic algorithms, 2023.

Tengyang Xie, Dylan Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed H. Awadallah, and Alexander Rakhlin. Exploratory Preference Optimization: Harnessing Implicit Q*-Approximation for Sample-Efficient RLHF. *International Conference on Learning Representations*, 2025:43632–43669, May 2025.

W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *ICML*, pp. 54715–54754, 2024.

H. Xu, A. Sharaf, Y. Chen, W. Tan, L. Shen, B. Van Durme, K. Murray, and Y. J. Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *ICML*, pp. 55204–55224, 2024a.

Mingjing Xu, Peizhong Ju, Jia Liu, and Haibo Yang. PSMGD: Periodic Stochastic Multi-Gradient Descent for Fast Multi-Objective Optimization, December 2024b.

S. Xu, W. Fu, J. Gao, W. Ye, W. Liu, Z. Mei, G. Wang, C. Yu, and Y. Wu. Is DPO superior to PPO for LLM alignment? A comprehensive study. In *ICML*, pp. 54983–54998, 2024c.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *ArXiv Preprint: 2505.09388*, 2025.

R. Yang, X. Pan, F. Luo, S. Qiu, H. Zhong, D. Yu, and J. Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *ICML*, pp. 56276–56297, 2024.

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

T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn. Gradient surgery for multi-task learning. In *NeurIPS*, pp. 5824–5836, 2020.

W. Yuan, R. Y. Pang, K. Cho, X. Li, S. Sukhbaatar, J. Xu, and J. E. Weston. Self-rewarding language models. In *ICML*, pp. 57905–57923, 2024.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *ArXiv Preprint: 2509.08827*, 2025.

Qi Zhang, Peiyao Xiao, Shaofeng Zou, and Kaiyi Ji. MGDA Converges under Generalized Smoothness, Provably. https://arxiv.org/abs/2405.19440v5, May 2024a.

Xiaoyuan Zhang, Xi Lin, and Qingfu Zhang. PMGDA: A Preference-based Multiple Gradient Descent Algorithm, February 2024b.

Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *ArXiv Preprint: 2305.10425*, 2023.

Shiji Zhou, Wenpeng Zhang, Jiyan Jiang, Wenliang Zhong, Jinjie Gu, and Wenwu Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond, 2022.

Z. Zhou, J. Liu, J. Shao, X. Yue, C. Yang, W. Ouyang, and Y. Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *ACL 2024 (Findings)*, pp. 10586–10613, 2024.

B. Zhu, M. I. Jordan, and J. Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *ICML*, pp. 43037–43067, 2023.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *ArXiv Preprint: 1909.08593*, 2019.

Appendix

A Further Related Work **19**

B Geometry and extensions of admissible preference sets **20**

B.1 Constructing ordinal preference sets 20

B.2 Admissible preference sets with general linear constraints 22

C Detail on Theoretical Analysis **24**

D Additional Experiment Setups and Results **24**

D.1 Performance 24

D.2 Ablation Studies 26

D.3 Missing Materials in Main Text 27

D.4 Case Studies 30

D.5 Experimental Overview 34

A Further Related Work

Preference learning. Learning from preference comparisons has become a central mechanism for aligning language models with human intent. RLHF learns reward models from comparisons and uses them to optimize language-model policies (Christiano et al., 2017; Ouyang et al., 2022), while Direct Preference Optimization (DPO) (Rafailov et al., 2023) directly optimizes the policy from preference pairs, avoiding a separate reward-model and RL stage. A broad family of subsequent methods modifies the preference objective, the available preference data, or the interaction protocol, developing alternative losses and feedback formulations (Ethayarajh et al., 2024; Tang et al., 2024; Meng et al., 2024; Liu et al., 2024a; Rafailov et al., 2024). Other work studies online preference collection and distribution mismatch between data and policies (Dong et al., 2024; Song et al., 2024; Shi et al., 2025). These methods primarily learn preferences between responses under a given alignment objective, whereas multi-objective alignment considers multiple objectives simultaneously and may additionally incorporate their relative importance (Wang et al., 2024a; Guo et al., 2024b; Wang et al., 2024b; Gupta et al., 2025; Liu et al., 2025). When multiple objectives are optimized jointly, gradient conflicts become a key challenge. Recently, Chen et al. (2026a) address such conflicts through conflict-aware modification of the update direction. However, many preference-aware multi-objective methods require numerical objective weights, which may be only partially known or subject to uncertainty. Our framework addresses this problem through an admissible preference set that combines ordinal constraints, optional cardinal weights, and a confidence parameter. This representation allows partially specified objective preferences to guide optimization without specifying a single fixed weight vector.

Multiobjective optimization methods. Multiobjective optimization studies problems in which several objectives must be optimized simultaneously and solutions are characterized through Pareto efficiency rather than a unique scalar optimum (Miettinen, 1999; Ehrgott, 2005). Classical approaches often transform the vector-valued problem into scalar subproblems through weighted-sum scalarization, ϵ-constraint method, normal boundary intersection, and normal constraint methods (Haimes et al., 1971; Das & Dennis, 1998; Messac & Mattson, 2004). Joint optimization of multiple objectives can lead to conflicting gradients, where descent directions for different objectives are misaligned. Gradient-based methods therefore construct common descent directions from the objective gradients, typically by solving an auxiliary problem to determine their combination (Jean-Antoine, 2012; Sener & Koltun, 2018; Fernando et al., 2022; Chen et al., 2023; Xiao et al., 2023; Zhang et al., 2024a). Preference-aware methods further steer optimization toward preferred regions of the Pareto set using reference vectors (Lin et al., 2019), explicit preference constraints (Zhang et al., 2024b), or preference functions over Pareto solutions (Chen et al., 2025a). These approaches primarily encode preference at the solution or Pareto-set level, rather than directly restricting the admissible combinations of objective gradients during optimization. Closely related to our formulation, Kaddani et al. (2017) represent preference through weight polytopes and use their extreme points to characterize and compute preferred nondominated solutions. However, a general polytope can be cumbersome for users to specify directly. Our formulation constructs the admissible set from interpretable cardinal, ordinal, and confidence information, providing a more structured and intuitive interface for preference specification. Moreover, by integrating this set, PECCO enforces preference consistency in gradient combinations throughout optimization, rather than using the preference set primarily to characterize or recover preferred nondominated solutions.

Multi-objective LLM alignment. Multi-objective LLM alignment methods manage competing behavioral criteria through model interpolation (Rame et al., 2023), preference-conditioned policies (Guo et al., 2024b; Wang et al., 2024b; Gupta et al., 2025), or adaptive objective weighting (Liu et al., 2025). Recent gradient-based approaches further address conflicts between objective gradients by modifying the update direction (Chen et al., 2026a). However, these methods generally do not ensure that the resulting gradient combinations remain consistent with a prescribed objective preference throughout training, and methods such as RACO require a numerical preference vector. Our work instead considers partially specified preferences and constrains gradient combinations to an admissible set defined by objective rankings, optional numerical weights, and confidence, allowing the weights to adapt to the gradients while preserving preference consistency.

B GEOMETRY AND EXTENSIONS OF ADMISSIBLE PREFERENCE SETS

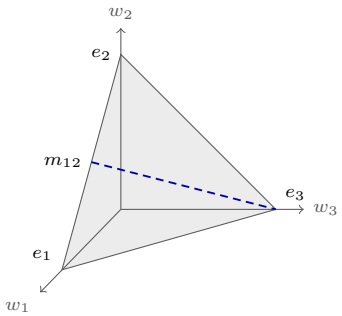
B.1 CONSTRUCTING ORDINAL PREFERENCE SETS

For three objectives $\mathcal{F} = \{f^1, f^2, f^3\}$, preference weights lie in

$$\Delta_3 = \{w \in \mathbb{R}^3 : w_i \geq 0,\ w_1 + w_2 + w_3 = 1\} = \mathrm{conv}\{e_1, e_2, e_3\},$$

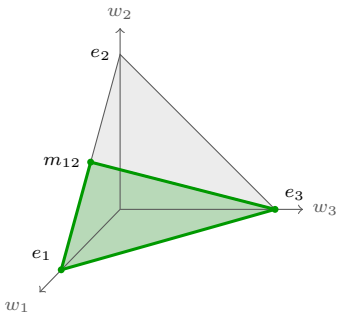
where e_i is the ith coordinate vector. This simplex is a triangle in three-dimensional weight space. All figures use the same view of this triangle.

The preference $f^1 \succeq f^2$ requires $w_1 \geq w_2$. To draw this set, connect the midpoint $m_{12} = (1/2, 1/2, 0)$ of the e_1e_2 edge to e_3. This segment is the boundary $w_1 = w_2$. Keep the half containing e_1, including the boundary (Figure 6). The result is $\mathcal{P}_{12} = \mathrm{conv}\{e_1, m_{12}, e_3\}$.



$w_1 = w_2$

(a) Draw the equality boundary



$\mathcal{P}_{12}:\ w_1 \geq w_2$

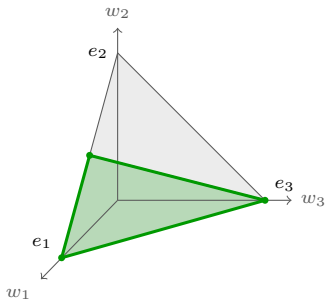
(b) Retain the half containing e_1

Figure 6: Drawing $w_1 \geq w_2$. Join m_{12} to e_3 to draw $w_1 = w_2$, then keep the half containing e_1. Green includes the boundary.

For several ordinal constraints, draw the set for each constraint and take their intersection. For example, to add $w_1 \geq w_3$, connect $m_{13} = (1/2, 0, 1/2)$ to e_2 and again keep the half containing e_1. Call this set $\mathcal{P}_{13}$. Its intersection with $\mathcal{P}_{12}$ is the quadrilateral in Figure 7:

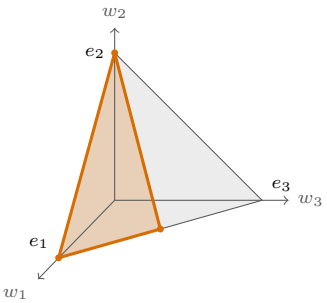
$$\mathcal{P}_{\mathrm{ord}} = \mathcal{P}_{12} \cap \mathcal{P}_{13} = \mathrm{conv}\{e_1, m_{12}, c, m_{13}\}, \qquad c = (1/3, 1/3, 1/3).$$

This construction applies to any specified preference relation $\succeq$. It gives the ordinal preference set defined in Section 4.1: $\mathcal{P}_{\mathrm{ord}} = \{w \in \Delta_3 : w_i \geq w_j \text{ whenever } f^i \succeq f^j\}$.



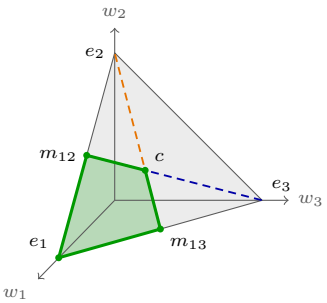
$\mathcal{P}_{12}:\ w_1 \geq w_2$

(a) First restriction



$\mathcal{P}_{13}:\ w_1 \geq w_3$

(b) Second restriction



$w_1 \geq w_2, \quad w_1 \geq w_3$

(c) Intersection

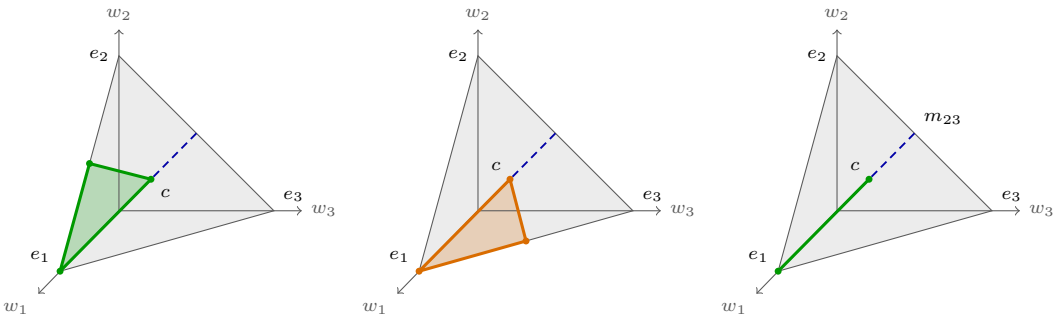
Figure 7: Adding $w_1 \geq w_3$. Panel (c) is the intersection of (a) and (b). The boundaries meet at $c = (1/3, 1/3, 1/3)$; the intersection has extreme points e_1, m_{12}, c, m_{13}.

To require equal weights, impose both opposite inequalities:

$$\{w : w_2 = w_3\} = \{w : w_2 \geq w_3\} \cap \{w : w_3 \geq w_2\}.$$

In the full simplex, $w_2 = w_3$ is the segment from e_1 to the midpoint $m_{23} = (0, 1/2, 1/2)$ of the e_2e_3 edge. To add this equality to the quadrilateral above, keep its portion satisfying $w_2 \geq w_3$ and its portion satisfying $w_3 \geq w_2$, then take their intersection (Figure 8). Only the segment from e_1 to c remains:

$$\mathcal{P}_{\text{ord}} \cap \{w : w_2 = w_3\} = \text{conv}\{e_1, c\} = \{(1 - 2t, t, t) : 0 \leq t \leq 1/3\}.$$



$\mathcal{P}_{\text{ord}} \cap \{w : w_2 \geq w_3\}$ $\mathcal{P}_{\text{ord}} \cap \{w : w_3 \geq w_2\}$ $\mathcal{P}_{\text{ord}} \cap \{w : w_2 = w_3\}$

(a) First weak inequality (b) Opposite weak inequality (c) Impose both: equality

Figure 8: Adding $w_2 = w_3$ to $\mathcal{P}_{\text{ord}} = \{w \in \Delta_3 : w_1 \geq w_2,\ w_1 \geq w_3\}$. Panels (a) and (b) impose the opposite inequalities separately. Their intersection in (c) is the segment $[e_1, c]$. The dashed extension to m_{23} shows the rest of the boundary $w_2 = w_3$ in the full simplex.

If no preference between f^2 and f^3 is specified, neither inequality is added. This is the incomparable case, denoted by $w_2 \parallel w_3$ in the main text. The full quadrilateral remains feasible, allowing $w_2 > w_3$, $w_3 > w_2$, and $w_2 = w_3$. Equal weights therefore impose an additional restriction; an unspecified ordinal relation does not.

B.2 ADMISSIBLE PREFERENCE SETS WITH GENERAL LINEAR CONSTRAINTS

Our extreme-point parameterization also works for general preference polytopes. Linear constraints specify a nonempty set $\mathcal{Q} = \{w \in \Delta_K : Aw \leq b,\ Ew = d\}$, its H-representation. Such sets can express partial preference information (Kaddani et al., 2017); Figure 9 shows twelve examples with three objectives. Converting the constraints to extreme points $p_1, \ldots, p_M$ gives the V-representation:

$$\mathcal{Q} = \mathrm{conv}\{p_1, \ldots, p_M\} = \{P_{\mathcal{Q}} u : u \in \Delta_M\}, \qquad P_{\mathcal{Q}} = [p_1, \ldots, p_M]. \tag{8}$$

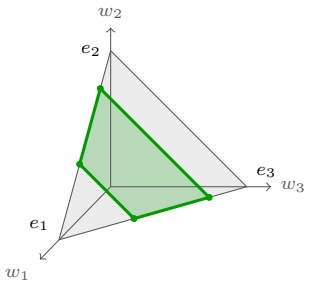
Here Δ_K and Δ_M are the probability simplices for K objectives and M extreme points. PECCO can then use $w = P_{\mathcal{Q}} u$ and the multi-gradient combination $\nabla F(\theta) P_{\mathcal{Q}} u$. Established conversion methods include primal–dual vertex enumeration (Bremner et al., 1998) and Qhull's halfspace intersection via geometric duality using Quickhull (Barber et al., 1996).

These constraints can also be combined with the ordinal, cardinal, and confidence specification in Section 4. First intersect $\mathcal{Q}$ with $\mathcal{P}_{\mathrm{ord}}$ to obtain a nonempty set $\mathcal{R} = \mathcal{Q} \cap \mathcal{P}_{\mathrm{ord}}$. For a numerical preference vector $\tilde{w} \in \mathcal{R}$, define

$$\mathcal{P}^{\mathcal{R}}_{\mathrm{adm}} = (1 - \lambda)\mathcal{R} + \lambda \tilde{w}, \qquad 0 \leq \lambda < 1. \tag{9}$$

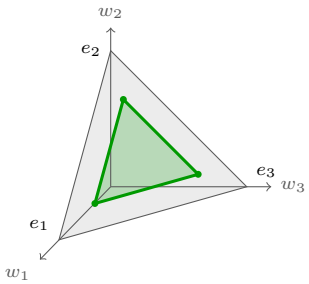
At $\lambda = 0$, the set is $\mathcal{R}$ and no numerical preference vector is needed. Increasing λ contracts it toward $\tilde{w}$ while preserving all constraints. Apply this same affine map to the extreme points of $\mathcal{R}$ to obtain the admissible matrix used by PECCO.

General linear constraints require users to choose coefficients and bounds and check that the resulting set is nonempty. They can also produce many extreme points, increasing conversion cost and the number of transformed gradients. We therefore use ordinal preference information, a numerical preference vector, and a confidence parameter in the main text: these inputs clearly express objective ordering, the desired numerical trade-off, and confidence in that trade-off. General polyhedral constraints remain an extension for more detailed preference information.



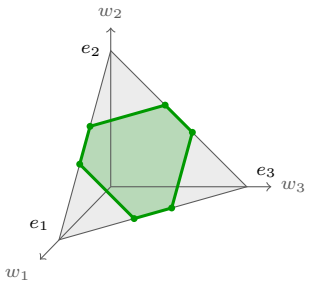
$0.2 \le w_1 \le 0.6$

(a) Weight interval



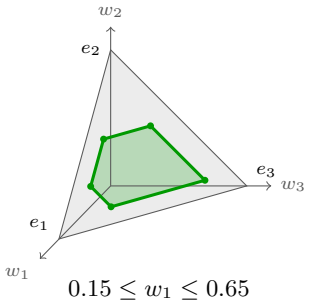
$w_i \ge 0.15 \quad (i = 1, 2, 3)$

(b) Minimum allocations



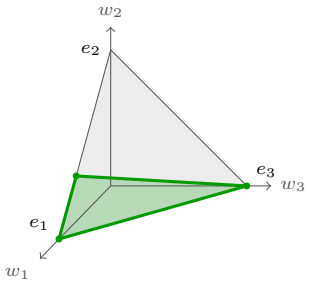
$w_i \le 0.6 \quad (i = 1, 2, 3)$

(c) Maximum allocations



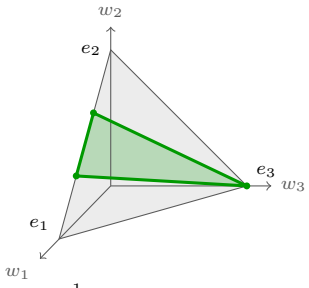
$0.15 \le w_1 \le 0.65$
$0.1 \le w_2 \le 0.5,\ w_3 \ge 0.1$

(d) Asymmetric intervals



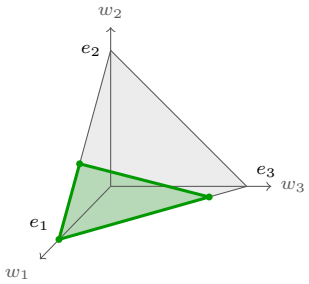
$w_1 \ge 2w_2$

(e) Relative-importance bound



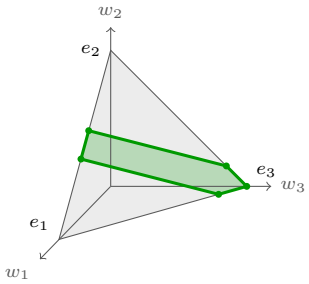
$\frac{1}{2}w_2 \le w_1 \le 2w_2$

(f) Relative-importance interval



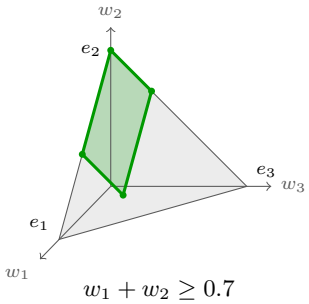
$w_1 - w_2 \ge 0.2$

(g) Minimum preference gap



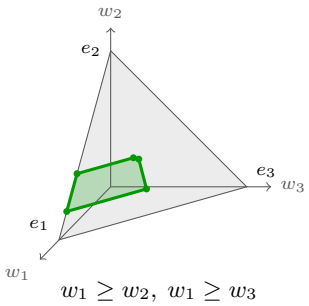
$|w_1 - w_2| \le 0.15$

(h) Approximate balance



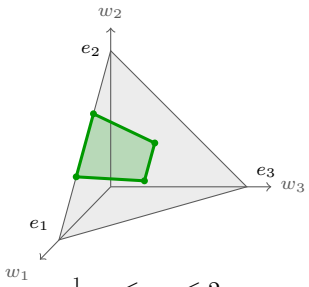
$w_1 + w_2 \ge 0.7$
$w_1 \le 0.55$

(i) Group-weight budget



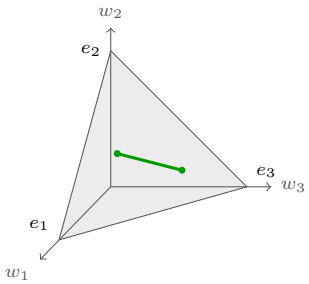
$w_1 \ge w_2,\ w_1 \ge w_3$
$0.15 \le w_2 \le 0.35$

(j) Ordinal with bounds



$\frac{1}{2}w_2 \le w_1 \le 2w_2$
$w_3 \le 0.4$

(k) Ratios with an allocation cap



$w_1 = w_2$
$0.2 \le w_3 \le 0.6$

(l) Equality with an interval

Figure 9: Twelve examples of linear constraints on preference weights. Each panel also requires $w \ge 0$ and $w_1 + w_2 + w_3 = 1$. Gray is the full simplex; green is $\mathcal{Q}$, which is the admissible preference set at $\lambda = 0$. Dots mark extreme points. Panels (a)–(k) are polygons in the simplex plane; (l) is a line segment. In (h), the absolute-value bound represents two linear inequalities. Each panel is a separate example.

C Detail on Theoretical Analysis

Let $H(\theta) = F(\theta)P_{\text{adm}} = (h^1(\theta), ..., h^M(\theta))$. Following the proof in Zhang et al. (2024a), we introduce Assumption C.1 of generalized smoothness and bounded function, and the corresponding Theorem C.2.

Assumption C.1 (Generalized smoothness and bounded value)**.** *The following assumptions are necessary for the convergence analysis:*

1. *For each $m \in [M]$, h^m is bounded by $h^{m,\star} := \inf_\theta h^m(\theta) > -\infty$.*
2. *For each $m \in [M]$, h^m are twice differentiable and generalized ℓ-smooth, i.e., $\|\nabla^2 h^m(\theta)\| \le \ell(\|\nabla h^m(\theta)\|)$ almost everywhere, where $\ell : [0, +\infty) \to (0, +\infty)$ is continuous and non-decreasing. Moreover, $\phi(a) := a^2/2\ell(2a)$ is monotonically increasing for any $a \ge 0$.*

Under these assumptions on H, we obtain:

Theorem C.2 (Theorem 1 in Zhang et al. (2024a))**.** *Suppose Assumption C.1 hold. Let $c > 0$ and $\mathcal{H} > 0$ be constants such that $\Delta + c \le \mathcal{H}$, where $\Delta := \max_{i\in[K]}\{h_i(\theta_0) - h_i^\star\}$, and define $\Gamma := \sup\{z \ge 0 : \phi(z) \le \mathcal{H}\}$. Set $\gamma = \mathcal{O}(1/\Gamma^2)$, $\alpha = \mathcal{O}(1/\Gamma^2 + 1/(\Gamma\ell(\Gamma + 1)))$, $T = \max\{\Theta(1/(\alpha\epsilon^2)), \Theta(1/(\gamma\epsilon^2))\}$, and $\rho = \mathcal{O}(\epsilon^2)$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla H(\theta_t)w_t\|^2 \le \epsilon^2.$$

Theorem 5.1 holds under the following assumptions on $f^k, \forall k \in [K]$.

Assumption C.3 ($L-$smoothness and bounded value)**.** *For each $k \in [K]$, f^k is lower bounded by $f^{k,\star} := \inf_\theta f^k(\theta) > -\infty$ and is L-smooth, i.e.,*

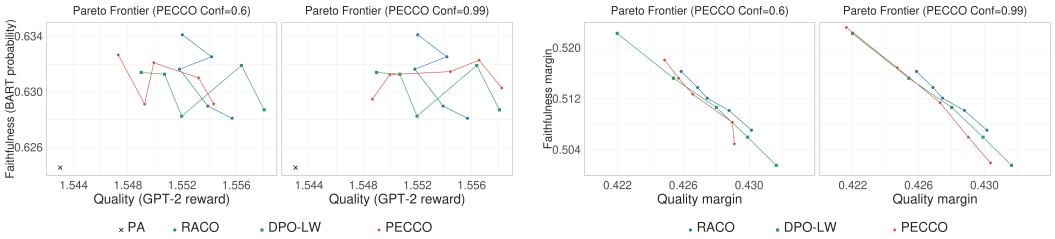
$$\|\nabla f^k(\theta) - \nabla f^k(\theta')\| \le L\|\theta - \theta'\|, \qquad \forall, \theta, \theta'.$$

This directly implies that functions $h^1, ..., h^M$ satisfy Assumption C.3 as well, which is a special case for Assumption C.1. Also, since $\nabla H(\theta_t) = \nabla F(\theta_t)P_{\text{adm}}$, applying Theorem C.2 on $h^1, ..., h^M$ under Assumption C.3 and taking $v_t = P_{\text{adm}}w_t$ immediately give us Theorem 5.1.

D Additional Experiment Setups and Results

D.1 Performance

Quality-faithfulness. Beyond the empirical analysis reported in the main text, we further evaluate PECCO on the *quality-faithfulness* task. We report the resulting Pareto trade-offs measured by judgment-model scores (Figure 10a) and preference margins (Figure 10b). The results are consistent with our previous observations, showing that PECCO remains competitive with the baselines. We use the same model and dataset as in the *quality-conciseness* analysis in the main text.



(a) Pareto frontiers (reward model). (b) Pareto frontiers (margin metric).

Figure 10: Pareto frontiers in *quality-faithfulness* task.

Gemma Models We repeat the performance test on *Gemma3-4B-Base* and *Gemma3-4B-Instruct* (Kamath et al., 2025) with the same setup as the *quality-conciseness* experiment in the main text. Results are shown in Figures 11, 12, and 13, and show similar performance trend.

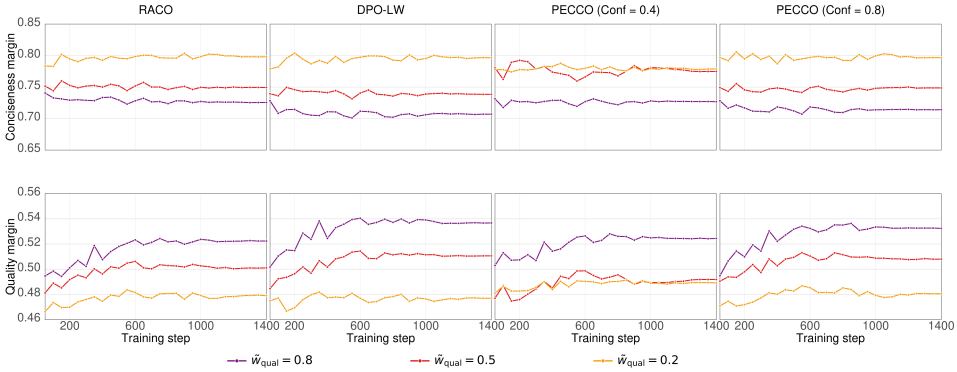


Figure 11: Trajectory on Gemma3-4B-Base

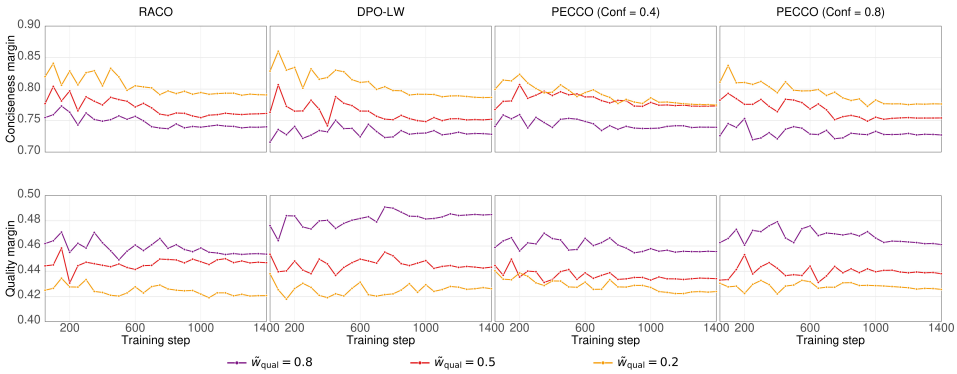


Figure 12: Trajectory on Gemma3-4B-Instruct

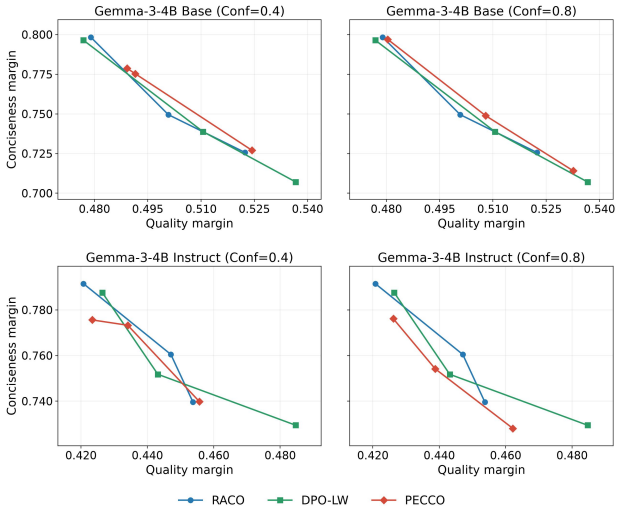


Figure 13: Pareto frontiers for Gemma3-4B models

Safety alignment (BeaverTails). We evaluate helpfulness–harmlessness alignment on Qwen3-4B-Instruct-2507 using PKU-SafeRLHF preference pairs and 700 prompts from BeaverTails-Evaluation (Ji et al., 2023). The training data provide separate helpfulness and harmlessness preference labels for response pairs. We use $w_{\text{help}} \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$, $w_{\text{safe}} = 1 - w_{\text{help}}$, and PECCO confidence levels $\{0.6, 0.8, 0.99\}$. Responses are generated with temperature 0.6 and at most 2048 new tokens. The official Beaver reward and cost models judge helpfulness and harmfulness, respectively. We compare responses to the same prompt at matching objective weights: higher reward indicates a helpfulness win, while lower cost indicates a harmlessness win. We report the two win rates separately, counting ties as half wins. Table 5 shows that PECCO achieves win rates above 50% in 14/15 helpfulness and 11/15 harmlessness comparisons against DPO-LW, and 14/15 and 10/15, respectively, against RACO. These results suggest modest improvements, particularly in helpfulness.

		Against DPO-LW		Against RACO	
Confidence	w_{help}	Helpfulness	Harmlessness	Helpfulness	Harmlessness
0.6	0.20	54.07	52.21	55.21	52.50
0.6	0.35	51.21	49.86	54.14	50.79
0.6	0.50	53.57	50.07	51.00	47.71
0.6	0.65	53.29	48.57	52.86	49.36
0.6	0.80	55.50	51.50	54.71	54.14
0.8	0.20	52.79	50.14	55.57	49.29
0.8	0.35	49.50	48.71	53.71	50.93
0.8	0.50	55.21	51.00	48.50	45.57
0.8	0.65	55.36	51.79	53.86	53.93
0.8	0.80	52.71	51.14	51.29	54.07
0.99	0.20	51.21	50.64	50.14	50.14
0.99	0.35	52.57	50.07	58.21	53.07
0.99	0.50	54.21	50.21	51.57	45.71
0.99	0.65	54.29	48.86	56.07	50.43
0.99	0.80	53.71	51.86	53.71	53.57

Table 5: PECCO win rates (%) against DPO-LW and RACO on BeaverTails. Comparisons use the same objective weights and 700 prompts, with Beaver reward/cost scores evaluating helpfulness and harmlessness separately; ties receive half credit. Across the 15 configurations, PECCO achieves win rates above 50% in 14/15 (helpfulness) and 11/15 (harmlessness) comparisons against DPO-LW, and 14/15 and 10/15, respectively, against RACO.

D.2 Ablation Studies

We conduct additional ablation experiments for PECCO on the parameters γ and ρ. Essential results are reported in Table 6. PECCO performs consistently across the tested ranges of γ and ρ, indicating limited sensitivity to these hyperparameters. The setting γ and $\rho = 3 \times 10^{-3}$ achieves the lowest validation loss and is therefore selected as the default configuration.

γ	ρ	Validation Loss ↓	Quality Margin ↑	Conciseness Margin ↑
3×10^{-5}	3×10^{-3}	0.836592	0.450614	0.753618
3×10^{-5}	3×10^{-2}	0.834470	0.449488	0.752693
3×10^{-5}	3×10^{-1}	0.837549	0.447815	0.754032
3×10^{-5}	3	0.835825	0.447931	0.752272
3×10^{-4}	3×10^{-3}	0.838219	0.449917	0.751864
3×10^{-4}	3×10^{-2}	0.838439	0.448992	0.752026
3×10^{-4}	3×10^{-1}	0.845004	0.448008	0.753694
3×10^{-4}	3	0.839019	0.447670	0.752822
3×10^{-3}	3×10^{-3}	0.833279	0.447042	0.754193
3×10^{-3}	3×10^{-2}	0.838837	0.448977	0.752469
3×10^{-3}	3×10^{-1}	0.837116	0.447968	0.754359
3×10^{-3}	3	0.833763	0.448842	0.753968

Table 6: Ablation over the iterative step size γ and regularization coefficient ρ for PECCO. All reported values are averaged over the final five validation evaluations.

D.3 Missing Materials in Main Text

We offer complete version of tables and figures for main text in this section.

Preference	Cardinal vector $\tilde{w}$	Confidence	Quality m_{qual}	Conciseness m_{conc}	Faithfulness m_{faith}
PA	–	–	0.4202	0.7429	0.4969
$f^1 \succeq f^2 \succeq f^3$	$(0.5, 0.3, 0.2)$	0.8	0.4333	0.7574	0.5091
$f^1 \succeq f^2 \succeq f^3$	$(0.6, 0.3, 0.1)$	0.8	0.4362	0.7548	0.5043
$f^1 \succeq f^2 \succeq f^3$	$(0.8, 0.15, 0.05)$	0.8	0.4410	0.7481	0.5017
$f^1 \succeq f^2 \succeq f^3$	–	–	0.4357	0.7513	0.5060
$f^1 \succeq f^3 \succeq f^2$	$(0.5, 0.2, 0.3)$	0.8	0.4343	0.7523	0.5134
$f^1 \succeq f^3 \succeq f^2$	$(0.6, 0.1, 0.3)$	0.8	0.4369	0.7476	0.5149
$f^1 \succeq f^3 \succeq f^2$	$(0.8, 0.05, 0.15)$	0.8	0.4444	0.7428	0.5079
$f^1 \succeq f^3 \succeq f^2$	–	–	0.4347	0.7533	0.5083
$f^2 \succeq f^1 \succeq f^3$	$(0.3, 0.5, 0.2)$	0.8	0.4266	0.7639	0.5106
$f^2 \succeq f^3 \succeq f^1$	$(0.2, 0.5, 0.3)$	0.8	0.4230	0.7664	0.5133
$f^3 \succeq f^1 \succeq f^2$	$(0.3, 0.2, 0.5)$	0.8	0.4282	0.7556	0.5264
$f^3 \succeq f^2 \succeq f^1$	$(0.2, 0.3, 0.5)$	0.8	0.4258	0.7589	0.5257
$f^1 \succeq \{f^2, f^3\},\ f^2 \parallel f^3$	$(0.5, 0.25, 0.25)$	0.8	0.4322	0.7566	0.5102
$f^1 \succeq \{f^2, f^3\},\ f^2 \parallel f^3$	$(0.6, 0.2, 0.2)$	0.8	0.4361	0.7517	0.5085
$f^1 \succeq \{f^2, f^3\},\ f^2 \parallel f^3$	$(0.8, 0.1, 0.1)$	0.8	0.4435	0.7446	0.5065
$f^1 \succeq \{f^2, f^3\},\ f^2 \parallel f^3$	–	–	0.4338	0.7534	0.5090
$f^2 \succeq \{f^1, f^3\},\ f^1 \parallel f^3$	$(0.25, 0.5, 0.25)$	0.8	0.4241	0.7657	0.5144
$f^2 \succeq \{f^1, f^3\},\ f^1 \parallel f^3$	$(0.2, 0.6, 0.2)$	0.8	0.4214	0.7680	0.5118
$f^2 \succeq \{f^1, f^3\},\ f^1 \parallel f^3$	$(0.1, 0.8, 0.1)$	0.8	0.4193	0.7770	0.5075
$f^2 \succeq \{f^1, f^3\},\ f^1 \parallel f^3$	–	–	0.4222	0.7698	0.5129

Table 7: PECCO under three-objective linear and partial-order preferences. Here, f^1, f^2, and f^3 denote quality, conciseness, and faithfulness, respectively. For experiments incorporating cardinal information, the confidence is fixed at $\lambda = 0.8$; "–" denotes preference-only information. PA denotes the pre-alignment policy.

Cardinal Preference	Confidence λ	Quality Margin m_{qual}	Conciseness Margin m_{conc}
$(0.2, 0.8)$	0.0	**0.4343**	0.7823
$(0.2, 0.8)$	0.2	0.4308	0.7878
$(0.2, 0.8)$	0.4	0.4280	0.7879
$(0.2, 0.8)$	0.6	0.4306	0.7901
$(0.2, 0.8)$	0.8	0.4282	**0.7928**
$(0.35, 0.65)$	0.0	**0.4364**	0.7844
$(0.35, 0.65)$	0.2	0.4328	0.7869
$(0.35, 0.65)$	0.4	0.4299	0.7860
$(0.35, 0.65)$	0.6	0.4300	**0.7885**
$(0.35, 0.65)$	0.8	0.4325	0.7823
$(0.45, 0.55)$	0.0	0.4318	0.7842
$(0.45, 0.55)$	0.2	0.4318	**0.7851**
$(0.45, 0.55)$	0.4	0.4335	0.7842
$(0.45, 0.55)$	0.6	0.4336	0.7803
$(0.45, 0.55)$	0.8	**0.4368**	0.7815
$(0.5, 0.5)$	0.0	0.4328	0.7847
$(0.5, 0.5)$	0.2	0.4325	**0.7848**
$(0.5, 0.5)$	0.4	0.4280	0.7847
$(0.5, 0.5)$	0.6	0.4324	0.7821
$(0.5, 0.5)$	0.8	**0.4386**	0.7719
$(0.55, 0.45)$	0.0	0.4455	0.7658
$(0.55, 0.45)$	0.2	0.4455	0.7685
$(0.55, 0.45)$	0.4	0.4426	0.7659
$(0.55, 0.45)$	0.6	0.4439	**0.7700**
$(0.55, 0.45)$	0.8	**0.4471**	0.7671
$(0.65, 0.35)$	0.0	0.4431	**0.7675**
$(0.65, 0.35)$	0.2	0.4448	0.7643
$(0.65, 0.35)$	0.4	0.4476	0.7622
$(0.65, 0.35)$	0.6	0.4482	0.7633
$(0.65, 0.35)$	0.8	**0.4509**	0.7613
$(0.8, 0.2)$	0.0	0.4425	**0.7692**
$(0.8, 0.2)$	0.2	0.4448	0.7641
$(0.8, 0.2)$	0.4	0.4509	0.7617
$(0.8, 0.2)$	0.6	0.4533	0.7571
$(0.8, 0.2)$	0.8	**0.4590**	0.7504

Table 8: Objective-specific validation margins at step 1400. Bold values indicate the best margin within each cardinal-preference group.

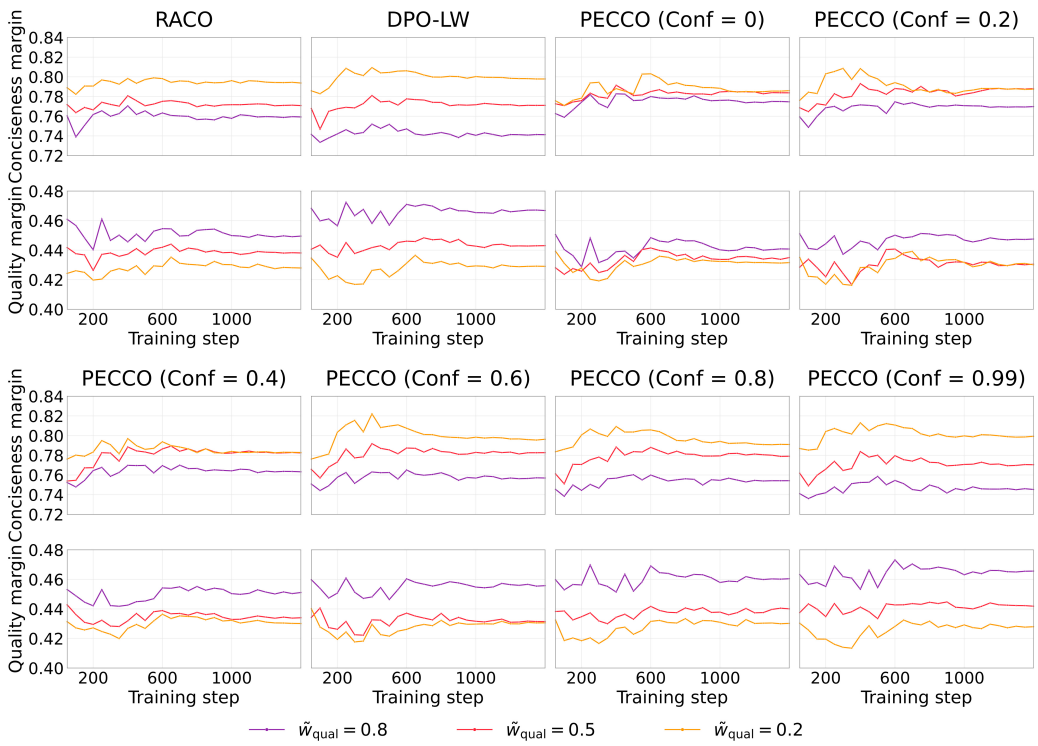


Figure 14: Complete performance trajectories

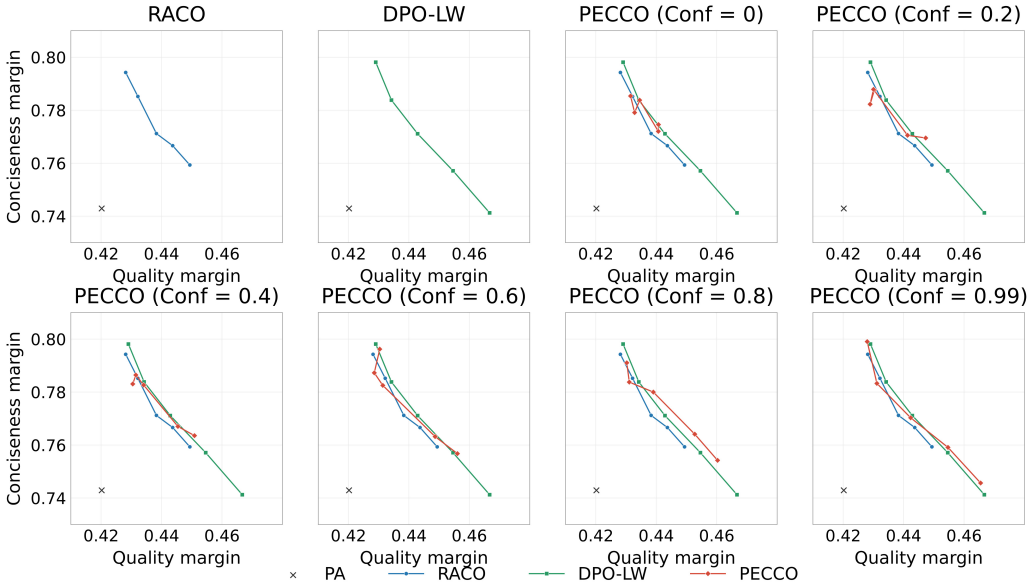


Figure 15: Complete performance Pareto frontier

D.4 CASE STUDIES

We examine generated summaries to make the quality–conciseness trade-off concrete for Qwen3-4B-Instruct-2507 and Llama-3.1-8B-Instruct. We compare PECCO under two preference configurations on Qwen and compare PECCO, RACO, and DPO-LW at a common configuration. These cases complement the validation preference metrics by showing which details are retained, omitted, or altered in the generated text.

Case-study protocol. We use the final exported checkpoints from the quality–conciseness experiments. Write the training weights as $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}})$, denoting quality and conciseness, respectively. PECCO's confidence is fixed at 0.6 throughout; the Qwen weight comparison uses $(0.8, 0.2)$ and $(0.65, 0.35)$, with the same ordinal priority qual $\succ$ conc in both configurations. Cross-method comparisons fix $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ and use the synchronized RACO and DPO-LW checkpoints.

Before generation, we fix 32 posts from the held-out TL;DR test set: the first 32 entries in our existing 2,000-post test file whose formatted prompts contain 100–260 whitespace-delimited words. Normalized post bodies do not overlap with the training or validation data. Each checkpoint receives the same posts and the instruction to produce a short paragraph. We use each backbone's chat template and generate one response per post with temperature 0.6, top-$p = 1$, seed 42, and a 512-token output limit. The $(0.65, 0.35)$ Qwen checkpoint was evaluated in a follow-up on these same posts after inspecting the initial weight comparison. Configurations and examples were selected exploratorily for qualitative analysis; they are not a random sample for estimating comparative performance. We reproduce complete source posts and all text returned by the generator, omitting only the common instruction wrapper. Source row indices are zero-based indices into the original test split. Word counts use whitespace splitting.

Example 1: More concise summaries with preserved central content. For Qwen3-4B-Instruct, the source describes a reader using *The New Bloomsday Book* alongside *Ulysses*: the guide improves appreciation of the novel but slows progress, threatening the reader's plans to finish other books before school resumes. With $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$, PECCO produces a 106-word summary that explicitly frames the reader's request as balancing comprehension and pace. Increasing the conciseness weight to 0.35, with $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.65, 0.35)$, yields an 83-word summary. The shorter response compresses the description and omits the final explicit advice-seeking sentence while retaining the guide's benefit, the slower progress, and the deadline concern. This selected pair illustrates the intended trade-off: less text while retaining the post's central dilemma.

Example 1: More concise summaries with preserved central content

Source Post: Verbatim; TL;DR test row 180 Subreddit: r/books Title: A query about Ulysses

I've finally gotten onto Ulysses after reading and loving Dubliners and Portrait, and I am definitely going to finish it, so that's not the issue for me.

However, I'm not quite sure how best to do this. I'm up to page 100 now, and I've been reading along with The New Bloomsday Book, alternating between reading a paragraph of the guide, then the corresponding paragraph of the actual novel.

The pros of this method: the guide is really making me appreciate the book more than I know I would without it; some of the commentary on the 'mn' and the gypsies and what have you makes me realise just how astonishingly written Ulysses really is, and I'm not sure that I'd be enjoying it anywhere near as much without the guide.

The con of this method: it's taking *forever.* I'm essentially reading a 250 page book and a 700 page book in tandem, which makes both take longer than they would otherwise. I really want to finish this book before the end of summer so I can cram in either the Brothers Karamazov or Infinite Jest before school starts again, but at this rate I don't think I'll be able to.

Output A: PECCO: PECCO, $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ (106 words). A Reddit user discusses their experience reading James Joyce's *Ulysses*, having started after enjoying *Dubliners* and *Portrait of the Artist as a Young Man*. They're using *The New Bloomsday Book* as a guide to enhance their understanding and appreciation of the novel, which significantly deepens their enjoyment but slows their progress due to reading both the novel and guide simultaneously. While the guide helps them appreciate Joyce's brilliance, the dual reading is consuming too much time, making it difficult to finish *Ulysses* before summer ends so they can move on to *Brothers Karamazov* or *Infinite Jest*. They're seeking advice on how to balance comprehension and pace.

Output B: PECCO: PECCO, $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.65, 0.35)$ (83 words). The user is reading James Joyce's *Ulysses* after enjoying *Dubliners* and *Portrait*, and is using *The New Bloomsday Book* as a guide to better understand and appreciate the novel. While the guide enhances their enjoyment by highlighting the book's complexity and brilliance, it significantly slows progress—reading both the novel and guide simultaneously makes the experience much longer. The user is concerned they won't finish *Ulysses* in time to read *Brothers Karamazov* or *Infinite Jest* before school starts, despite their enthusiasm for the book.

Example 2: Preserving event order across methods. We next compare the three methods on Qwen3-4B-Instruct, holding $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ fixed. The source contains two successive mishaps: the vacuum cylinder first spills onto the floor, and the vacuum subsequently catches and overturns the trash bag during cleanup. RACO compresses these into an initial spill caused by the bag being caught and adds an unsupported expansion of the acronym TIFU. DPO-LW and PECCO preserve the two-stage sequence without that expansion. Their responses contain 88 and 85 words, respectively, compared with RACO's 79. The relevant distinction is preservation of the causal sequence rather than word count alone; this example supports both DPO-LW and PECCO over RACO on that specific aspect.

Example 2: Preserving event order across methods

Source Post: Verbatim; TL;DR test row 17 Subreddit: r/tifu Title: TIFU by emptying the vacuum
So my wife and I are adopting two kittens today and have been preparing the kitten room, cleaning thoroughly and setting everything up. Vacuuming was my job, and I was doing a damn good job (IMO). I should inform you that we have one of those Dyson Cyclone vacuums that has the cylinder that you detach and then press the button to empty from the bottom. So I detach the cylinder and hold it over the trash bag to empty. I press the button-dust, dirt and hair ALL OVER the floor while only a small amount made it into the bag. I curse silently, turn the vacuum back on and go to work. I had gotten most of it back into the vacuum when the garbage bag was caught by the vacuum and tipped over, spilling the rest of the vacuum's prior contents onto the floor. I cursed again (not silently this time) and cleaned the rest of it up. Such a simple task and yet, TIFU.

RACO Output: RACO, $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ (79 words). The author shares a humorous "TIFU" (Today I Felt Unfortunate) story about accidentally spilling vacuum contents during a cleaning prep for adopting kittens. While vacuuming with a Dyson Cyclone, they detached the cylinder to empty it into a trash bag, only to have dust, dirt, and hair scatter across the floor due to the bag getting caught in the vacuum's mechanism. After recovering the mess, they realized how easily a simple task went wrong—highlighting the absurdity of the situation.

DPO-LW Output: DPO-LW, $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ (88 words). The user shares a TIFU about accidentally spilling vacuum contents during a cleaning task for their new kitten room. While vacuuming, they detached the Dyson Cyclone cylinder to empty it into a trash bag, only to have dust, dirt, and hair scatter across the floor. After retrieving most of the debris, the trash bag got caught in the vacuum and tipped over, releasing the remaining contents onto the floor. The incident, though simple, led to a frustrated cleanup and highlighted how easily a small mistake can go wrong.

PECCO Output: PECCO, $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ (85 words). The user recounts a TIFU incident while preparing a kitten room for adoption. While vacuuming with a Dyson Cyclone, they detached the cylinder to empty it into a trash bag, only to have dust, dirt, and hair spill out onto the floor. After briefly cleaning it up, they realized the trash bag had been caught by the vacuum and tipped over, spilling the remaining debris. The experience, though simple, turned into a frustrating and messy situation, highlighting how even routine tasks can go wrong unexpectedly.

Example 3: Factual restraint and omitted communicative intent. On Llama-3.1-8B-Instruct, the source describes an unreturned bass guitar after a breakup. RACO and PECCO each produce 75-word responses that retain the unreturned instrument, failed meetings, and blocked communication. DPO-LW produces 121 words, introduces the unsupported year 2021, and misrepresents the source's timing as an ex-boyfriend 'of 2 weeks.' PECCO avoids those additions, but it also leaves the author's explicit request for advice implicit, whereas RACO retains it. Thus, PECCO is more restrained than DPO-LW in this example, while RACO better preserves the post's communicative purpose. The comparison does not establish a uniform winner across all aspects of summarization.

Example 3: Factual restraint and omitted communicative intent

Source Post: Verbatim; TL;DR test row 330 Subreddit: r/relationships Title: I [19F] lent my bass guitar to my now ex-boyfriend [22M]. I'd like it back, but he's ignoring me.
A couple weeks before we broke up, I lent my now ex-boyfriend my bass guitar so he could record a song (it's a nice bass and he's always loved playing it). He lives two hours away from me, so after the break-up we planned on meeting before one of his weekly choir practices, which is half way between us, so he could give it back.
Three months later, I don't have it. We've planned on meeting many times, but he's cancelled each one. Now, I haven't even heard from him in weeks—he hasn't answered my calls, he's blocked me on Facebook, he doesn't log into Skype, and didn't answer me when I messaged him via his music page on Facebook. Instead, he posted a status about all the cool bass lines he's been writing.
As you can imagine, I'm quite frustrated. I know he had a hard time with us breaking up, but it's been months and he's still acting like a kid. I've never been in this position; what's the appropriate way to handle this? Threaten to get police involved? Call his mom? I don't want to be a crazy ex-girlfriend.

RACO Output: RACO, $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ (75 words). I lent my bass guitar to my ex-boyfriend of 3 months ago for him to record a song, but he's since broken up with me, and he's not returning the instrument, despite our planned meetings and promises to return it. He's blocked me on social media, so I'm unable to reach him, but he's posting about his music on Facebook, but not to me. I'm frustrated and looking for advice on how to handle this.

DPO-LW Output: DPO-LW, $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ (121 words). I lent my bass guitar to my ex-boyfriend of 2 weeks in 2021, but 3 months later, I haven't received it back. We broke up and we set a meeting schedule, but he's been ignoring me when it's time to return my instrument. I've been unable to reach him via phone, Facebook, or other ways. We're in a large area of the country, and it's 3 months later, yet he's still avoiding me for the return of my bass. I'm not sure what to do, as I'm not sure how to handle this situation. We've been broken up, but he's still acting as if he's someone who's not willing to follow through on his promises, and this is leaving me frustrated.

PECCO Output: PECCO, $(\tilde{w}_{\text{qual}}, \tilde{w}_{\text{conc}}) = (0.8, 0.2)$ (75 words). I lent my bass guitar to my ex-boyfriend for recording a song in a couple weeks before our breakup, with the understanding he'd return it later. We broke up, and now three months later he's ignoring me altogether, leaving me with no way to retrieve my instrument. Despite our planned meetings, he's consistently cancelled, and I've tried to contact him through various channels, but he's blocked me, and it's as if he's intentionally avoiding me.

Interpretation. These cases separate three aspects of a summary: coverage of source details, fidelity to the source, and the amount of generated text. A shorter response can omit useful context, while a longer response can add either relevant detail or unsupported elaboration. The validation preference metric measures relative likelihood on fixed candidate pairs; it does not directly measure the length or factual accuracy of newly sampled summaries. Consequently, these examples illustrate behavior at specific configurations, rather than establishing a monotonic length response to the weights or a uniform ranking of methods. The study uses one training seed and one generation per post; statistical comparisons require a broader evaluation.

D.5 Experimental Overview

We organize the experiments by their purpose in Table 9. The configuration tables retain only the model, methods, data, and essential hyperparameters. We write γ for the PECCO weight-update step size and β for the preference-loss parameter.

Table 9: Overview of experiments and configuration tables.

Experiment	Purpose	Setup
Performance	Compare objective trade-offs across tasks and models.	Table 10
Preference consistency	Examine training weights and preference-dependent outcomes.	Table 11
Gradient conflict avoidance	Measure conflicts in the combined update direction.	Table 12
Confidence parameter ablation	Examine how confidence changes the trade-off.	Table 13
Algorithm hyperparameter ablation	Assess sensitivity to γ and ρ.	Table 14
Qualitative case studies	Inspect generated summaries and their limitations.	Table 15

Table 10: Performance across summarization and safety tasks.

Item	Configuration / description
Experiment	Performance: quality–conciseness, quality–faithfulness, and helpfulness–harmlessness.
Purpose and result	Compare objective trade-offs. PECCO is competitive across tasks and backbones; BeaverTails comparisons show modest gains, particularly in helpfulness.
Models	Qwen3-4B-Instruct-2507; Llama-3.1-8B-Instruct; Gemma-3-4B-Base and Gemma-3-4B-Instruct. Llama and Gemma are used for quality–conciseness; the other tasks use Qwen.
Methods	RACO, DPO-LW, PECCO.
Datasets	RACOo/RedditSummary-Alignment for summarization, with BART-derived faithfulness labels for quality–faithfulness; PKU-SafeRLHF for safety training and BeaverTails-Evaluation for testing.
Shared hyperparameters	Global batch size 64; maximum sequence length 2,048; $\beta = 0.2$. One epoch with linear learning-rate decay. Initial learning rate: 2×10^{-5} for Qwen/Llama quality–conciseness and safety; 1.5×10^{-5} for Gemma; 2×10^{-6} for quality–faithfulness.
RACO hyperparameters	Default settings (Chen et al., 2026a).
PECCO hyperparameters	Qwen/Gemma quality–conciseness: $\gamma = 3 \times 10^{-5}$, $\rho = 3 \times 10^{-3}$. Llama, quality–faithfulness, and safety use the closed-form solver; γ and ρ are not used.

Table 11: Preference consistency in training weights and alignment outcomes.

Item	Configuration / description
Experiment	Preference consistency.
Purpose and result	Inspect training weights and three-objective alignment outcomes. PECCO weights remain in the admissible region, while RACO weights can depart from it. PECCO yields distinct trade-offs consistent with the supplied preferences.
Model	Qwen3-4B-Instruct-2507
Methods	RACO and PECCO for training-weight distributions; PECCO for three-objective alignment outcomes.
Dataset	RACOo/RedditSummary-Alignment with BART-derived faithfulness labels.
Shared hyperparameters	Global batch size 64; maximum sequence length 2,048; $\beta = 0.2$. Initial learning rate 10^{-5}; linear decay; one epoch.
RACO hyperparameters	Default settings (Chen et al., 2026a).
PECCO hyperparameters	$\gamma = 3 \times 10^{-3}$; $\rho = 3 \times 10^{-3}$.

Table 12: Gradient conflict avoidance during quality–conciseness training.

Item	Configuration / description
Experiment	Gradient conflict avoidance.
Purpose and result	Measure conflicts between the combined update and the reference gradients. PECCO has near-zero conflict rates with its admissible transformed gradients; baseline rates use the original objective gradients.
Model	Qwen3-4B-Instruct-2507
Methods	RACO, DPO-LW, PECCO.
Dataset	RACOo/RedditSummary-Alignment
Shared hyperparameters	Global batch size 64; maximum sequence length 2,048; $\beta = 0.2$. Initial learning rate 2×10^{-5}; linear decay; one epoch.
RACO hyperparameters	Default settings (Chen et al., 2026a).
PECCO hyperparameters	$\gamma = 3 \times 10^{-5}$; $\rho = 3 \times 10^{-3}$.

Table 13: Ablation on the confidence parameter.

Item	Configuration / description
Experiment	Ablation on confidence parameter λ.
Purpose and result	Examine how confidence in numerical preferences affects quality–conciseness alignment. Varying confidence changes the resulting objective trade-offs.
Model	Qwen3-4B-Instruct-2507
Methods	PECCO; RACO and DPO-LW provide fixed references in the comparison plots.
Dataset	RACOo/RedditSummary-Alignment
Shared hyperparameters	Global batch size 64; maximum sequence length 2,048; $\beta = 0.2$. Initial learning rate 2×10^{-5}; linear decay; one epoch.
RACO hyperparameters	Default settings (Chen et al., 2026a).
PECCO hyperparameters	Confidence-sweep plots: $\gamma = 3 \times 10^{-5}$, $\rho = 3 \times 10^{-3}$. The step-1,400 margin tables use the closed-form solver, without γ or ρ.

Table 14: Ablation on PECCO algorithm hyperparameters.

Item	Configuration / description
Experiment	Ablation on the weight-update step size γ and regularization coefficient ρ.
Purpose and result	Assess sensitivity on quality–conciseness alignment. Validation losses and margins vary little across the tested grid; $\gamma = \rho = 3 \times 10^{-3}$ gives the lowest validation loss.
Model	Qwen3-4B-Instruct-2507
Methods	PECCO.
Dataset	RACOo/RedditSummary-Alignment
Shared hyperparameters	Global batch size 64; maximum sequence length 2,048; $\beta = 0.2$. Initial learning rate 2×10^{-5}; linear decay; 700 training updates.
RACO hyperparameters	Not applicable.
PECCO hyperparameters	$\gamma \in \{3 \times 10^{-5}, 3 \times 10^{-4}, 3 \times 10^{-3}\}$; $\rho \in \{3 \times 10^{-3}, 3 \times 10^{-2}, 3 \times 10^{-1}, 3\}$.

Table 15: Qualitative comparison of generated summaries.

Item	Configuration / description
Experiment	TL;DR case studies.
Purpose and result	Inspect coverage, factuality, and conciseness in generated summaries. The examples expose omissions and unsupported details without establishing a uniform ranking of methods.
Models	Qwen3-4B-Instruct-2507 and Llama-3.1-8B-Instruct.
Methods	RACO, DPO-LW, PECCO.
Dataset	32 held-out TL;DR posts; existing quality–conciseness checkpoints.
Shared hyperparameters	No additional training. Checkpoints: initial learning rate 2×10^{-5}; global batch size 64; one epoch; maximum sequence length 2,048; $\beta = 0.2$. Generation: temperature 0.6; at most 512 new tokens.
RACO hyperparameters	Default settings (Chen et al., 2026a).
PECCO hyperparameters	Closed-form solver; γ and ρ are not used.