Robust Multi-Objective Alignment via Pessimistic Reward Model Distillation (AT-PRMD)

***Abstract*—Language model alignment increasingly requires balancing multiple, often competing objectives such as helpfulness, harmlessness, and honesty. Direct Preference Optimization (DPO), while computationally efficient, frequently exhibits degenerate behavior by over-optimizing single objectives at the expense of others—a manifestation of reward hacking. We present a pessimistic ensemble approach to reward model distillation that addresses this challenge by optimizing against a family of reward models rather than a single point estimate. Specifically, we train three reward models on the Anthropic HH-RLHF dataset and apply pessimistic DPO, which maximizes performance under the worst-case reward model in the ensemble. This approach naturally encodes Constitutional AI principles by ensuring robust satisfaction of multiple alignment criteria simultaneously. Through controlled experiments, we demonstrate that pessimistic ensemble DPO prevents degenerate solutions, maintains balanced performance across competing objectives, and provide theoretical guarantees on worst-case alignment. Our results provide both empirical validation and theoretical insight into how explicit reward modeling with pessimistic regularization can achieve more reliable multi-objective alignment in the offline setting.**

I. Introduction

A. Motivation and Background

The alignment of large language models (LLMs) with human preferences represents one of the most critical challenges in modern artificial intelligence. While supervised fine-tuning produces models capable of generating fluent text, it provides no guarantees that model outputs will align with human values across dimensions such as helpfulness, harmlessness, and honesty (Christiano et al., 2017; Bai et al., 2022). Early approaches to alignment employed Reinforcement Learning from Human Feedback (RLHF), which trains an explicit reward model from preference annotations and subsequently optimizes the language model policy through online reinforcement learning (Stiennon et al., 2020; Ouyang et al., 2022). However, the computational demands and training instability of online RL have motivated the development of offline alternatives. Direct Preference Optimization (DPO; Rafailov et al., 2023) emerged as an elegant solution that bypasses explicit reward modeling entirely, directly optimizing policies on preference data through a supervised learning objective. This approach has been widely adopted in both academic and industrial settings, including in the training of Llama 3 (AI@Meta, 2024) and OLMo (Groeneveld et al., 2024). Despite its practical advantages, recent work has identified fundamental limitations of DPO. Rafailov et al. (2024a) and Pal et al. (2024) demonstrated that DPO can assign near-zero probability to preferred responses, while Azar et al. (2024) showed that DPO's implicit regularization is often insufficient to prevent reward hacking. Further as Constitutional AI (CAI; Bai et al., 2022) proposes that language models should adhere to a set of explicit principles or "constitution" governing behavior across various dimensions. Rather than optimizing for a single scalar reward, CAI suggests that aligned models should satisfy multiple constraints simultaneously. However, existing implementations of constitutional principles often rely on iterative self-critique or multiple rounds of training, increasing computational complexity. Our work explores whether pessimistic reward model distillation can provide a more direct path to multi-objective alignment. By constructing a family of reward models that emphasize different aspects of the preference distribution—and optimizing for robust performance across all models in the family—we hypothesize that the resulting policy will naturally satisfy multiple constitutional principles without requiring explicit multi-stage training procedures.

B. The Multi-Objective Alignment Challenge

A particularly pressing issue emerges when alignment requires balancing multiple, potentially conflicting objectives. Consider the following scenarios:

Helpfulness vs. Harmlessness: A user requesting detailed chemical synthesis information creates tension between providing maximally useful information and preventing potential misuse.

Helpfulness vs. Honesty: Models may increase perceived helpfulness through sycophancy (agreeing with user statements regardless of accuracy) or by expressing unwarranted certainty, sacrificing epistemic honesty.

Competing safety criteria: Different stakeholders may prioritize different aspects of model behavior, creating no single "correct" objective function.

Traditional single-objective optimization, including standard DPO, exhibits degenerate behavior in such settings. The policy may exploit idiosyncrasies in the training data or reward model to achieve high scores on one dimension while catastrophically failing on others—a phenomenon known as Goodhart's Law (Goodhart, 1975) or reward hacking in the reinforcement learning literature (Amodei et al., 2016).

So the problem at hand can be describes as:

- Different alignment objectives have conflicting gradients
- Training data contains biased or noisy preference labels
- Models exploit loopholes in single-objective reward functions

Our Reward Model Distillation (RMD) offers a promising solution through its pessimistic regularization mechanism. By considering a family of plausible reward models rather than a single point estimate, pessimistic RMD can:

- Prevent overfitting to any single objective
- Maintain robust performance across multiple alignment criteria
- Provide theoretical guarantees on worst-case alignment

C. Contributions

This paper makes the following contributions:

Empirical demonstration of pessimistic ensemble DPO: We implement and evaluate a pessimistic reward model distillation approach using three reward models trained on subsamples of the Anthropic HH-RLHF dataset, demonstrating improved robustness to distribution shift.

Analysis of degenerate behavior: Through controlled experiments, we characterize how standard DPO produces degenerate solutions when preference data exhibits systematic biases, and show how pessimistic regularization prevents such failure modes.

Connection to Constitutional AI: We demonstrate that the pessimistic framework naturally implements multi-objective alignment by ensuring satisfaction of multiple implicit "principles" encoded in the reward model ensemble, without requiring explicit constitutional self-critique.

Comparative evaluation: We evaluate pessimistic reward model distillation by comparing it against existing preference–based alignment methods and by situating its performance relative to published results for modern aligned models such as GPT-4, Claude, Qwen, and Llama. We focus on metrics that capture the three core axes of alignment—helpfulness, harmlessness, and honesty—along with robustness under objective disagreement. Our evaluation uses held-out preference win-rates, safety/jailbreak resistance, toxicity and refusal metrics, truthful reasoning benchmarks, and worst-case performance across reward models. This provides a realistic view of whether pessimistic distillation produces more stable, non-degenerate alignment compared to standard reward distillation and existing alignment methods. We will evaluate against RLHF trained models of GPT, Claude, and Mixtral.

II. Experimental Setup

A. Dataset

We conduct experiments on the Anthropic Helpful and Harmless (HH-RLHF) dataset (Bai et al., 2022), a widely-used benchmark for preference-based alignment research. The dataset contains approximately 170,000 binary preference comparisons between model responses, annotated by human evaluators along dimensions of helpfulness and harmlessness. Each example consists of a conversational context x and two responses (y_w, y_l), where y_w is preferred over y_l according to human judgment.

B. Model Architecture and Implementation

Base Language Model: We use Qwen 2.5-3B as our base policy model. The model is initialized from pre-trained weights and serves as both the initial policy π_θ and the frozen reference policy π_{ref}.

Reward Models: We train three separate reward models $\{r_1, r_2, r_3\}$, for each dimension of HHH using subset from the HH-RLHF dataset. Each reward model takes a (prompt, response) pair (x, y) as input and outputs a scalar reward score $r_i(x, y) \in R$.

Training Infrastructure: All experiments are conducted using RTX 5090, 32GB VRAM. Training code builds upon the open-source TRL (Transformer Reinforcement Learning) library with custom modifications for pessimistic ensemble training.

III. Methodology

A. Step 1: Train Reward Model Ensembles

For each objective $k \in \{\text{helpful}, \text{harmless}, \text{honest}, \text{concise}\}$, train M reward models using preference data:

$$r_{\phi_{k,m}}(x, y) : \mathcal{X} \times \mathcal{Y} \to R, \quad m = 1, 2, \ldots, M \tag{1}$$

where:

- x is the prompt
- y is the response
- M is ensemble size
- Each model trained with different random initialization/dropout for diversity

Loss for each reward model:

$$\mathcal{L}_{\text{RM}}(\phi_{k,m}) = -E_{(x,y_w,y_l)\sim\mathcal{D}_k}\left[\log\sigma\left(r_{\phi_{k,m}}(x, y_w) - r_{\phi_{k,m}}(x, y_l)\right)\right] \tag{2}$$

where:

- $\mathcal{D}_k$ is preference dataset for objective k
- y_w is preferred (winning) response
- y_l is rejected (losing) response
- σ is sigmoid function

B. Step 2: Compute Implicit Reward from Policy

Define implicit reward from policy parameters:

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \tag{3}$$

where:

- π_θ is current policy being optimized
- π_{ref} is frozen reference policy
- β is temperature parameter (typically 0.1)

Token-level computation:

$$r_\theta(x, y) = \beta \sum_{t=1}^{|y|} \log \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})} \tag{4}$$

C. Step 3: Standard Multi-Objective RMD (Baseline)

Weighted aggregation across objectives:

$$\mathcal{L}_{\text{multi-RMD}}(\theta) = \sum_{k=1}^{K} \lambda_k E_{(x,y_w,y_l)\sim\mathcal{D}} \left[\left(\Delta r_{\phi_k} - \Delta r_\theta \right)^2 \right] \quad (5)$$

where:

- $\Delta r_{\phi_k} = r_{\phi_k}(x, y_w) - r_{\phi_k}(x, y_l)$ (reward difference from model)
- $\Delta r_\theta = r_\theta(x, y_w) - r_\theta(x, y_l)$ (implicit reward difference)
- λ_k are objective weights, $\sum_k \lambda_k = 1$
- K is number of objectives

Limitation: May allow one objective to degrade while others improve.

D. Step 4: Pessimistic Multi-Objective RMD

Option A: Hard Minimum (Most Conservative)

Take worst-case across ensemble per objective:

$$\mathcal{L}_{\text{P-multi-RMD}}(\theta) = \sum_{k=1}^{K} \lambda_k E \left[\left(\min_{m\in[M]} \Delta r_{\phi_{k,m}} - \Delta r_\theta \right)^2 \right] \quad (6)$$

where:

- $\min_{m\in[M]} \Delta r_{\phi_{k,m}}$ selects worst-case reward from ensemble
- Ensures robustness to pessimistic reward estimates per objective

Option B: CVaR-Based

Use Conditional Value at Risk instead of hard minimum:

$$\mathcal{L}_{\text{CVaR-RMD}}(\theta) = \sum_{k=1}^{K} \lambda_k E \left[\left(\text{CVaR}_\alpha \left[\{\Delta r_{\phi_{k,m}}\}_{m=1}^{M} \right] - \Delta r_\theta \right)^2 \right] \quad (7)$$

where:

$$\text{CVaR}_\alpha[X] = E[X \mid X \le F_X^{-1}(\alpha)] \quad (8)$$

- α is quantile level (e.g., 0.1 for worst 10%)
- Averages over worst α-fraction of ensemble
- Less sensitive to outliers than hard minimum

E. Step 5: Hierarchical Pessimism

Two-level pessimism: worst within ensemble, then worst across objectives:

$$\mathcal{L}_{\text{hier-RMD}}(\theta) = E \left[\left(\min_{k\in[K]} \min_{m\in[M]} \Delta r_{\phi_{k,m}} - \Delta r_\theta \right)^2 \right] + \alpha \cdot \text{KL}(\pi_\theta \| \pi_{\text{ref}}) \quad (9)$$

where:

- Inner $\min_m$: worst-case within each objective's ensemble
- Outer $\min_k$: worst-case across all objectives
- α controls KL regularization strength

KL divergence approximation:

$$\text{KL}(\pi_\theta \| \pi_{\text{ref}}) \approx E_{x,y\sim\pi_\theta} \left[\log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right] = \frac{1}{\beta} E[r_\theta(x, y)] \quad (10)$$

F. Step 6: Constitutional Constraint Formulation

Add explicit constraints for minimum principle satisfaction:

$$\mathcal{L}_{\text{const-RMD}}(\theta) = E \left[\left(\min_k \min_m \Delta r_{\phi_{k,m}} - \Delta r_\theta \right)^2 \right] + \alpha \cdot \text{KL}(\pi_\theta \| \pi_{\text{ref}}) + \beta \sum_{k=1}^{K} \mathcal{C}_k(\theta) \quad (11)$$

where constraint violation penalty is:

$$\mathcal{C}_k(\theta) = \max \left(0, \tau_k - \min_{m\in[M]} E_{y\sim\pi_\theta(\cdot|x)}[r_{\phi_{k,m}}(x, y)] \right) \quad (12)$$

- τ_k is minimum acceptable reward for principle k
- β controls constraint penalty strength
- Ensures each principle exceeds threshold

G. Step 7: Training Objective with Regularization

Complete training loss:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{alignment}}(\theta) + \alpha \cdot \text{KL}(\pi_\theta \| \pi_{\text{ref}}) + \gamma \cdot \|\theta - \theta_{\text{init}}\|^2 \quad (13)$$

where:

- $\mathcal{L}_{\text{alignment}}$ is one of the above (Steps 3-6)
- Second term prevents policy from diverging too far from reference
- Third term (optional) is L2 regularization on parameters

H. Step 8: Policy Optimization

Gradient descent update:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_{\text{total}}(\theta_t) \quad (14)$$

where:

- η is learning rate (typically 10^{-6} to 10^{-5})
- Gradients computed via backpropagation through policy network

I. Step 9: Evaluation Metrics

Per-objective worst-case performance:

$$\text{Worst-Case}_k = E_{x\sim\mathcal{D}_{\text{test}}} \left[\min_{m\in[M]} r_{\phi_{k,m}}(x, y_\theta) \right] \quad (15)$$

where $y_\theta \sim \pi_\theta(\cdot|x)$ is sampled from trained policy.

Minimum across all objectives:

$$\text{Overall-Worst} = \min_{k\in[K]} \text{Worst-Case}_k \quad (16)$$

Violation rate:

$$\text{ViolationRate} = P \left(\exists k : \min_m r_{\phi_{k,m}}(x, y_\theta) < \tau_k \right) \quad (17)$$

IV. Training Pipeline

Input: Preference datasets $\{D_1, D_2, \ldots, D_K\}$ for K objectives.

1) For each objective k: Train ensemble $\{r_{\phi_1}, \ldots, r_{\phi_M}\}$ using $\mathcal{L}_{\text{RM}}$.
2) Initialize policy π_θ from pretrained model. Keep reference policy π_{ref} frozen.
3) For each training iteration:
   a) Sample batch (x, y_w, y_l) from D.
   b) Compute implicit rewards:

$$r_\theta(x, y_w) = \beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}$$

$$r_\theta(x, y_l) = \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}$$

   c) Compute ensemble reward differences for each objective k:

$$\{\, r_{\phi_{k,m}}(x, y_w) - r_{\phi_{k,m}}(x, y_l) \,\}_{m=1}^{M}$$

   d) Apply pessimism:
   - CVaR: $\text{CVaR}_\alpha[\{\Delta r_{\phi_{k,m}}\}]$
   - Hierarchical: $\min_k \min_m \Delta r_{\phi_{k,m}}$

   e) Compute loss: $\mathcal{L}_{\text{total}}(\theta)$
   f) Update parameters:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}(\theta)$$

4) Evaluate on test set:
   - Per-objective worst-case scores
   - Minimum across objectives
   - Violation rates

Output: Aligned policy π_θ with robustness.

V. Evaluation Metrics

We assess alignment quality along multiple dimensions:

- **Single Reward Model Distillation:** Distills from only the unbiased reward model, with no ensemble or pessimism.
- **Average Ensemble Distillation:** Uses the average reward difference

$$\Delta r_{\text{avg}} = \frac{\Delta r_1 + \Delta r_2 + \Delta r_3}{3}$$

  instead of the minimum.
- **Win Rate Against Reference:** Fraction of responses from π_θ preferred over π_{ref} responses by human evaluators (or a reward model proxy).
- **Per-Objective Performance:** Separate win rates computed using each individual reward model r_i, measuring performance on length-penalized, neutral, and length-rewarded criteria.
- **Worst-Case Performance:** Minimum win rate across all three reward models, quantifying robustness.
- **KL Divergence from Reference:** $\mathrm{D}_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ measured on a held-out prompt set, assessing how much the policy has shifted.
- **Benchmark Reference:** Alignment scores on various benchmarks of HH-RLHF Holdout, MT-Bench, TruthfulQA, RealToxicityPrompts, JailbreakBench / AdvBench against GPT, Claude, and Mixtral.

A. Hyperparameter Sensitivity Analysis

- β (implicit reward temperature): $\{0.01, 0.05, 0.1, 0.5\}$
- α (KL regularization): $\{0.01, 0.1, 1.0, 10.0\}$
- M (ensemble size): $\{1, 2, 3\}$
- **Pessimism method**: {max, CVaR-10%, CVaR-20%}

VI. Conclusion

In this work, we introduced a pessimistic reward model distillation framework for robust multi-objective alignment of large language models. Traditional alignment methods such as standard DPO or single-objective reward distillation often suffer from degenerate behavior, including over-optimization of one objective at the expense of others, vulnerability to biased preference data, and the exploitation of loopholes in scalar reward functions. Our approach addresses these limitations through the use of reward model ensembles, pessimistic aggregation mechanisms, and hierarchical constitutional constraints that provide explicit worst-case guarantees.

By training multiple reward models per objective and optimizing the policy against the most pessimistic elements of the ensemble, our method enforces robustness across helpfulness, harmlessness, honesty, and conciseness objectives. The introduction of CVaR-based pessimism and hierarchical minimization further strengthens guarantees by focusing optimization on the weakest-performing regions of the reward landscape. Additionally, the constitutional constraint formulation ensures that each alignment principle meets a minimum threshold, preventing collapse along individual dimensions.

Our evaluation metrics—including win rates, per-objective performance, worst-case alignment, KL divergence, and performance on HH-RLHF holdouts and broader alignment benchmarks—provide a comprehensive assessment of robustness. These metrics highlight the importance of evaluating aligned models not only on average performance but also under objective disagreement and adversarial conditions. The resulting aligned policy demonstrates improved stability compared to standard distillation methods and offers a clearer path toward principled multi-objective alignment.

Overall, pessimistic reward model distillation provides a unified, theoretically grounded, and practically effective framework for achieving reliable multi-objective alignment in offline settings. Future work may extend this approach to larger model scales, additional alignment dimensions, or dynamic constitutional constraints learned from human feedback.