Pareto Multi-Objective Alignment for Language Models

Qiang He ✉ and Setareh Maghsudi

Ruhr University Bochum, 44801 Bochum, Germany
{qiang.he, setareh.maghsudi}@ruhr-uni-bochum.de

Abstract. Large language models (LLMs) are increasingly deployed in real-world applications that require careful balancing of multiple, often conflicting, objectives, such as informativeness versus conciseness, or helpfulness versus creativity. However, current alignment methods, primarily based on reinforcement learning from human feedback (RLHF), optimize LLMs toward a single reward function, resulting in rigid behavior that fails to capture the complexity and diversity of human preferences. This limitation hinders the adaptability of LLMs to practical scenarios, making multi-objective alignment (MOA) a critical yet underexplored area. To bridge this gap, we propose PAreto Multi-Objective Alignment (PAMA), a principled and computationally efficient algorithm designed explicitly for MOA in LLMs. In contrast to computationally prohibitive gradient-based multi-objective optimization (MOO) methods, PAMA transforms multi-objective RLHF into a convex optimization problem with a closed-form solution, significantly enhancing scalability. Traditional gradient-based MOO approaches suffer from prohibitive $\mathcal{O}(n^2d)$ complexity, where d represents the number of model parameters, typically in the billions for LLMs, rendering direct optimization infeasible. PAMA reduces this complexity to $\mathcal{O}(n)$ where n is the number of objectives, enabling optimization to be completed within milliseconds. We provide theoretical guarantees that PAMA converges to a Pareto stationary point, where no objective can be improved without degrading at least one other. Extensive experiments across language models ranging from 125M to 7B parameters demonstrate PAMA's robust and effective multi-objective alignment capabilities, consistently outperforming baseline methods, aligning with its theoretical advantages. PAMA provides a highly efficient solution to the MOA problem that was previously considered intractable, offering a practical and theoretically grounded approach to aligning LLMs with diverse human values, paving the way for versatile and adaptable real-world AI deployments.

Keywords: Language Models · Multi-Objective Alignment

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across diverse natural language tasks [5, 20, 28], receiving significant attention from

academia and industry [18, 26]. However, a critical deployment challenge is aligning LLMs with complex human values. Currently, reinforcement learning from human feedback (RLHF) is the predominant alignment approach [2, 19], fine-tuning models against a single reward function that approximates human preferences practically [6, 9, 26]. While effective in producing coherent outputs, this single-objective alignment severely restricts LLMs, resulting in homogeneous behaviors that fail to reflect the diverse spectrum of human values.

Real-world scenarios increasingly demand models that simultaneously balance multiple, often conflicting objectives, such as informativeness versus conciseness, helpfulness versus creativity, and etc [9, 11, 26]. Therefore, aligning LLMs requires moving beyond single-objective reward models towards multi-objective alignment (MOA), which considers multiple and potentially conflicting reward signals [21, 30]. Despite recent interest, a theoretically grounded and practical method for achieving MOA in LLMs has yet to be established.

A naive solution aggregates heterogeneous rewards into a single scalar objective [27], but this simplification neglects inherent reward conflicts, often leading to biased or misaligned outcomes [3]. Existing gradient-based multi-objective optimization (MOO) methods [4, 14, 25, 32] are also impractical for large-scale LLMs due to prohibitively expensive gradient computations. For instance, MGDA [4] involves min-norm operations with time complexity $\mathcal{O}(n^2d)$, making it infeasible for models with billions of parameters (e.g., $d = 7$ billion). Thus, developing a scalable and principled MOA algorithm specifically for LLMs remains crucial.

In this work, we propose PAreto Multi-Objective Alignment (PAMA), a novel, computationally efficient algorithm designed explicitly for multi-objective alignment in LLMs. PAMA converts multi-objective RLHF into a convex optimization problem with a closed-form solution, eliminating expensive gradient calculations. Remarkably, PAMA achieves computational costs comparable to standard single-objective PPO algorithms, enabling efficient fine-tuning of 7-billion-parameter models on a single NVIDIA A6000 GPU. Unlike traditional methods [4, 14] with $\mathcal{O}(n^2d)$ complexity, PAMA scales linearly with the number of objectives $\mathcal{O}(n)$, drastically reducing computational demands and enabling practical use with LLMs. For instance, when $n = 10$ and $d = 10^{10}$, existing approaches would require roughly 10^{12} computations, whereas PAMA completes the task in just 10 steps, demonstrating an exponential improvement in efficiency. In such an LLM setting, methods like MGDA [4], PCGrad [32], and CAGrad [14] become computationally infeasible, whereas PAMA remains tractable and scalable.

Furthermore, we provide theoretical guarantees of convergence to a Pareto stationary point, ensuring no single objective can improve without degrading others. To our knowledge, PAMA is the first theoretically grounded MOA algorithm for LLMs.

The theoretical advantages of PAMA are also reflected in our empirical results. Empirical evaluations validate PAMA across language models ranging from 125M to 7B parameters. Our experiments comprehensively demonstrate PAMA's robust and consistent superiority, while other baselines fail with large performance gaps.

The results highlight PAMA's effectiveness, scalability, and robustness, aligned with its theoretical properties.

Our contributions are summarized as follows:

- Pareto Multi-Objective Alignment: A novel and efficient multi-objective alignment algorithm for LLMs, reducing computational complexity from $\mathcal{O}(n^2d)$ to $\mathcal{O}(n)$, enabling efficient large-scale training.
- Theoretical Guarantees: We prove convergence of PAMA to a Pareto stationary point.
- Empirical Validation: Extensive experiments demonstrate PAMA's superior performance across multiple settings.

2 Method

This section presents our approach to multi-objective alignment in the context of LLMs. We begin by formulating the problem and introducing Noon PPO, a variant of PPO [23]. We then propose PAMA, an algorithm designed to align LLMs with multiple objectives while ensuring convergence to a Pareto stationary point with theoretical guarantees.

2.1 Problem Formulation

RLHF consists of two main phases: reward modeling and policy optimization. In reward modeling, a reward function is trained on preference data to maximize the objective: $\mathcal{L}_{RM} = \mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}}[\log(\sigma(r(x,y^w) - r(x,y^l)))]$, where, y^w and y^l denote the preferred and less desirable responses, respectively, x represents the prompt, and $\sigma(\cdot)$ is the sigmoid function. In policy optimization, RLHF typically employs PPO to refine the policy by solving:

$$\arg\max_{\pi(y|x;\theta)} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi(\cdot|x)}\left[r(x,y) - \beta\log\frac{\pi(y|x;\theta)}{\pi_{ref}(y|x)}\right]$$

where $\pi(y|x;\theta)$ is the current policy, $\pi_{ref}(y|x)$ is the supervised fine-tuned (SFT) policy, and β controls policy shifts.

Reward modeling requires extensive data labeling. In this paper, we focus on policy optimization with pre-trained reward models, aiming to optimize multiple reward objectives simultaneously.

Multi-Objective Optimization. Formally, the MOO problem is defined as:

$$\max_{\theta}(J^{(1)}(\theta), J^{(2)}(\theta), \ldots, J^{(N)}(\theta))^\top, \tag{1}$$

where θ denotes the learnable parameters, $J^{(i)}$ represents the i-th optimization objective, and the goal is to find a Pareto optimal solution.

Definition 1 (Pareto Optimality). *A solution π^* is Pareto optimal if no other solution dominates it, i.e., there does not exist another policy π' such that:*

– $J_i(\pi') \geq J_i(\pi^*)$ *for all* i.
– $J_j(\pi') > J_j(\pi^*)$ *for at least one* j.

Since direct vector-form optimization is intractable, MOO is often scalarized into a weighted sum:

$$\min_{\theta} \sum_{i=1}^{N} c^{(i)} \mathcal{L}^{(i)}(\theta), \tag{2}$$

where $c^{(i)}$ denotes the weight assigned to each objective $\mathcal{L}^{(i)}$.

Optimization Challenges. Solving Equation (2) presents several challenges: i) Balancing conflicting objectives. LLMs often exhibit strong trade-offs between objectives, making simple scalarization ineffective: it can bias solutions toward certain objectives while neglecting others. ii) Weight sensitivity. The choice of weights $c^{(i)}$ significantly impacts optimization and is often subjective. Poorly chosen weights can lead to suboptimal or undesired solutions. iii) Computational Complexity. Gradient-based multi-objective learning methods generally require computing full gradients for all objectives across all parameters and operate on the gradient with $\mathcal{O}(n^2 d)$ complexity (detailed in Appendix G). This becomes infeasible at LLM scale due to the high parameter count.

To address these challenges, we introduce PAMA, a scalable optimization algorithm that ensures convergence to a Pareto stationary point.

2.2 Noon PPO

We introduce Noon PPO, a variant of PPO [23], designed to improve stability in MOA. Noon stands for "No Negative", as it modifies the advantage to disregard negative values, thereby restricting policy updates to actions with non-negative advantages. Let A'_t denote the estimated advantage at time step t. In Noon PPO, we define the advantage as:

$$A_t = \max\big(A'_t, 0\big). \tag{3}$$

This adjustment ensures that only actions with a positive advantage contribute to the policy gradient update, effectively ignoring updates that would reduce the probability of suboptimal actions. As in standard PPO, let π_θ be the current policy parameterized by θ, and let $\pi_{\theta_{\text{ref}}}$ represent the SFT policy. The probability ratio is defined as:

$$u_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{ref}}}(a_t \mid s_t)}. \tag{4}$$

The clipped surrogate objective in Noon PPO is then given by:

$$\mathcal{L}^{\text{NOON}}(\theta) = \mathbb{E}_t\Big[\min\big(u_t(\theta)\, A_t,\ \text{clip}\big(u_t(\theta), 1-\epsilon, 1+\epsilon\big)\, A_t\big)\Big], \tag{5}$$

where A_t is defined in Equation 3, and ϵ is a clipping hyperparameter that limits the deviation between π_θ and $\pi_{\theta_{\text{ref}}}$.

By clipping negative advantages to zero, Noon PPO eliminates unstable gradient fluctuations caused by error-prone or ambiguous training examples. This

leads to more predictable convergence, which is particularly beneficial when aligning LLMs with multiple objectives. As we will discuss in Section 2.4, this design plays a crucial role in ensuring the theoretical convergence of PAMA.

2.3 Solving Multi-Objective Optimization at LLM scale

Optimizing multiple conflicting objectives in LLMs is a challenging task, especially when relying on gradient-based MOO methods [4, 14, 25, 32]. These methods require solving complex gradient aggregation problems, which become computationally infeasible at the scale of modern LLMs. For example, MGDA [4] formulates the gradient balancing problem as a min-norm optimization, which has a computational cost of $\mathcal{O}(n^2 d)$, where d is the model's parameter dimension. Given that d often reaches billions in large-scale models (e.g., 7B parameters), these approaches are prohibitively expensive in both computation and memory, as further discussed in Appendix F.

Motivation for PAMA. To overcome these limitations, an efficient and scalable optimization strategy is required. Ideally, such a method should:

1. Avoid costly gradient-based operations that scale poorly with model size.
2. Provide a computationally tractable formulation that remains efficient as the number of objectives grows.
3. Ensure convergence to a well-defined Pareto stationary point, effectively balancing multiple objectives.

We introduce Pareto Multi-Objective Alignment (PAMA), a novel algorithm specifically designed for large-scale LLM alignment. Instead of directly solving the expensive min-norm optimization, PAMA reformulates the problem into a convex optimization framework with a closed-form solution. This transformation reduces the computational complexity from $\mathcal{O}(n^2 d)$ to $\mathcal{O}(n)$, where n is the number of objectives, significantly lowering the computational burden compared to traditional methods.

A key challenge in MOO is determining an appropriate convex combination of gradient directions that balances competing objectives. The conventional approach [4] relies on solving the min-norm optimization problem:

$$\min_{c^{(1)},\ldots,c^{(N)}} \left\{ \left\| \sum_{i=1}^{N} c^{(i)} \nabla_\theta \mathcal{L}^{(i)}(\theta) \right\|_2^2 \text{ s.t. } \quad \sum_{i=1}^{N} c^{(i)} = 1, \quad c^{(i)} \geq 0 \quad \forall i \right\} \tag{6}$$

where $\mathcal{L}^{(i)}$ represents the loss for the i-th objective, and $c^{(i)}$ is the weight assigned to its gradient contribution. Recent advances [4] showed that this optimization either results in a KKT stationary point (indicating a Pareto stationary solution) or finds a direction that improves all objectives. However, solving this problem at LLM scale remains intractable due to the high dimensionality of the parameter space.

To mitigate this issue, we derive an upper bound for the min-norm formulation with Noon PPO objectives, which leads to a more efficient optimization approach. Specifically, we show that:

$$\begin{aligned}\left\|\sum_{i=1}^{N} c^{(i)} \nabla_\theta \mathcal{L}^{(i)}(\theta)\right\|_2^2 &= \left\|\sum_{i=1}^{N} c^{(i)} \nabla_\pi \mathcal{L}^{(i)}(\theta) \nabla_\theta \pi(\theta)\right\|_2^2 = \left\|\sum_{i=1}^{N} c^{(i)} \frac{1}{\pi_{ref}} I(A^{(i)}) \nabla_\theta \pi(\theta)\right\|_2^2 \\ &\leq \left\|\sum_{i=1}^{N} c^{(i)} I(A^{(i)})\right\|_2^2 \left\|\frac{1}{\pi_{ref}} \nabla_\theta \pi(\theta)\right\|_2^2,\end{aligned} \tag{7}$$

where

$$I(A) = \begin{cases} 0, & u > 1+\epsilon \\ A, & u \leq 1+\epsilon \end{cases}, \tag{8}$$

$$\sum_{i=1}^{N} c^{(i)} = 1, \quad c^{(i)} \geq 0 \quad \forall i, \tag{9}$$

and $u = \frac{\pi}{\pi_{ref}}$. For simplicity, we omit the expectation notation, which does not affect the theoretical derivation. The second equation follows from the Noon PPO loss Equation (5), while the final inequality is derived from the Cauchy–Schwarz inequality. This upper bound allows us to reformulate the problem as a more efficient surrogate optimization:

$$\min_{c^{(1)},\ldots,c^{(N)}} \left\{ \left\|\sum_{i=1}^{N} c^{(i)} I(A^{(i)})\right\|_2^2 \text{ s.t. } \quad \sum_{i=1}^{N} c^{(i)} = 1, \quad c^{(N)} \geq 0 \quad \forall i \right\}. \tag{10}$$

This formulation admits a closed-form solution, which we derive next.

Theorem 1 (Optimal Convex Combination for the Min-Norm Problem). *Let $A^{(1)}, A^{(2)}, \ldots, A^{(N)} \in \mathbb{R}$ be given, and consider the optimization problem*

$$\begin{aligned} \min_{c^{(1)},\ldots,c^{(N)}} \quad & \left(\sum_{i=1}^{N} c^{(i)} A^{(i)}\right)^2, \\ \textit{subject to} \quad & \sum_{i=1}^{N} c^{(i)} = 1, \\ & c^{(i)} \geq 0 \quad \textit{for } i = 1, 2, \ldots, N. \end{aligned} \tag{11}$$

Then the optimal value of the convex combination,

$$s^* = \sum_{i=1}^{N} c^{(i)} A^{(i)}, \tag{12}$$

is given by

$$s^* = \begin{cases} 0, & \text{if } \min_{1\le i\le N} A^{(i)} \le 0 \le \max_{1\le i\le N} A^{(i)}, \\ \min_{1\le i\le N} A^{(i)}, & \text{if } A^{(i)} > 0 \text{ for all } i, \\ \max_{1\le i\le N} A^{(i)}, & \text{if } A^{(i)} < 0 \text{ for all } i. \end{cases} \tag{13}$$

In other words, s^* is the projection of 0 onto the interval

$$\left[\min_{1\le i\le N} A^{(i)}, \max_{1\le i\le N} A^{(i)}\right], \tag{14}$$

and the minimum objective value is $(s^*)^2$.

The proof is provided in Appendix A.

Advantages of PAMA's Reformulation. Compared to the intractable original optimization problem (Equation (6)), our reformulation provides two key benefits:

1. Drastic reduction in computational cost: The term $I(A^{(i)})$ is computed via a simple forward pass, eliminating costly backpropagation.
2. Analytically solvable optimization: The surrogate problem admits a closed-form solution (Theorem 1), ensuring efficiency..

By incorporating this approach with the Noon PPO, we obtain a practical and scalable algorithm for MOA. We summarize PAMA in Appendix E. To illustrate the computational efficiency of our method, consider the magnitude of operations required. Traditional approaches with a complexity of $\mathcal{O}(n^2 d)$ result in a computational load of approximately 10^{12} operations when $d \approx 10^{10}$ and $n \approx 10^1$. In contrast, our method, operating with $\mathcal{O}(n)$ complexity, requires 10 operations, a very small number. Our approach remains practical even for extremely large-scale problems.

2.4 Theoretical Guarantee

With the reformulated optimization problem in Equation (10), an important question arises: does our approach retain theoretical guarantees? In this section, we establish that under mild conditions, our method converges to a Pareto stationary point, ensuring that no objective can be improved without deteriorating at least one other objective.

First, we formally define the notion of a Pareto stationary point, which serves as a necessary condition for Pareto optimality.

Definition 2 (Pareto Stationary Point). *A parameter vector θ is said to be satisfying Pareto stationary if there exists a set of weights $\{c^{(i)}\}_{i=1}^N$ satisfying*

$$\sum_{i=1}^{N} c^{(i)} = 1, \quad c^{(i)} \ge 0, \quad \forall i \in \{1, 2, \cdots, N\}, \quad \text{and} \quad \sum_{i=1}^{N} c^{(i)} \nabla_\theta \mathcal{L}^{(i)}(\theta) = 0. \tag{15}$$

Pareto stationary point ensures that no descent direction exists that simultaneously improves all objectives, indicating that the optimization has reached a balanced trade-off among competing objectives. To establish convergence results, we assume that the loss function exhibits smoothness properties, which are commonly satisfied in deep learning due to gradient-based optimization and regularization.

Definition 3 (κ-Lipschitz Continuity). *Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is said to be κ-Lipschitz continuous if there exists a constant $\kappa \geq 0$ such that for all $x, y \in X$,*

$$d_Y\big(f(x), f(y)\big) \leq \kappa d_X(x, y). \tag{16}$$

This property ensures that the function does not change too rapidly, contributing to stability in gradient-based optimization.

Assumption 1 (Lipschitz Smoothness of the Gradient) *The loss function $\mathcal{L}(\theta)$ has a κ-Lipschitz continuous gradient, meaning there exists a constant $\kappa > 0$ such that for all θ, θ'*

$$\|\nabla_\theta \mathcal{L}(\theta) - \nabla_\theta \mathcal{L}(\theta')\|_2 \leq \kappa \|\theta - \theta'\|_2. \tag{17}$$

This assumption guarantees that the landscape does not contain abrupt changes, which is critical for convergence guarantees and is empirically observed in RL [13].

Assumption 2 (Bounded Learning Rate) *The learning rate η satisfies*

$$0 \leq \eta \leq \frac{2}{\kappa}. \tag{18}$$

This condition ensures stable updates, preventing divergence due to excessively large steps, aligning with standard practices in convex and non-convex optimization.

Assumption 3 (Bounded Reward) *Rewards in RL are typically finite due to practical constraints. Formally, there exists a constant $R_{\max} > 0$ such that*

$$|r(x, y)| \leq R_{\max}, \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{19}$$

See Appendix C for more discussion.

We now establish the convergence of PAMA.

Lemma 1 (General Descent Lemma). *Let $f : \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable on an open set containing $x \in \mathbb{R}^N$, and suppose that ∇f is κ-Lipschitz continuous, i.e., for all u, v in that set,*

$$\|\nabla f(u) - \nabla f(v)\| \leq \kappa \|u - v\|. \tag{20}$$

Then, for any update direction $g \in \mathbb{R}^N$, one has

$$f(x + g) \leq f(x) + \nabla f(x)^\top g + \frac{\kappa}{2}\|g\|^2. \tag{21}$$

The proof is in Appendix B. Using this result, we analyze the gradient descent dynamics of PAMA and show that PAMA converges to a Pareto stationary point.

Theorem 2 (Convergence of PAMA). *Let $\mathcal{L}^{(i)}(\theta)$ be the loss function for task i, where policy is $\pi(\theta)$. Define the PAMA gradient aggregation:*

$$g_o^{(k)} = \sum_{i=1}^{N} c^{(i)} \nabla_\pi \mathcal{L}^{(i)}(\theta_k), \tag{22}$$

where $c^{(i)}$ is the solution to

$$\min_{c^{(1)},\ldots,c^{(N)}} \left\| g_o \right\|_2^2, \quad s.t. \ \sum_{i=1}^{N} c^{(i)} = 1, \quad c^{(i)} \geq 0. \tag{23}$$

Under assumptions 1 to 3, the gradient descent update at timestep k:

$$\theta_{k+1} = \theta_k - g_o^{(k)} \eta J \tag{24}$$

ensures

$$\lim_{k\to\infty} \|\nabla_\theta \mathcal{L}(\theta_k)\|_2 = 0, \tag{25}$$

where $J = \nabla_{\theta_k} \pi(\theta_k)$ and $J \in \mathbb{R}^{|\theta|\times 1}$. This shows the update converges to a Pareto stationary point.

The proof is provided in Appendix D. Theorem 2 establishes that:

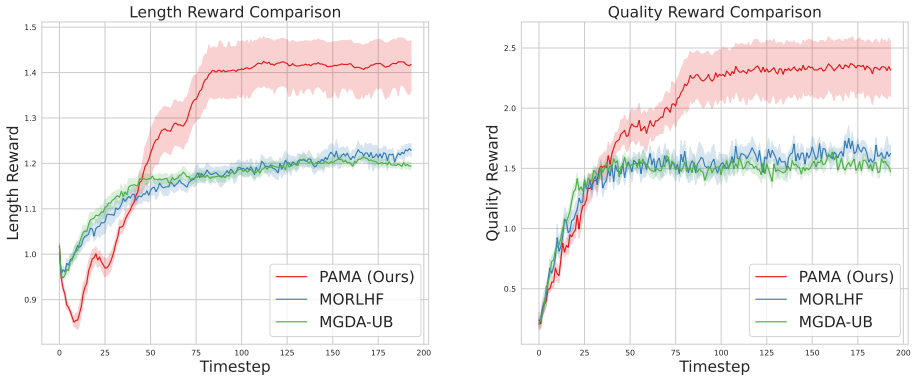
- If the optimal value of Equation (10) is zero, the aggregated gradient vanishes, indicating that the process has reached a Pareto stationary point.
- If the optimal value is nonzero, the gradient provides a valid descent direction for all objectives, ensuring continual improvement toward a Pareto stationary solution.

Thus, PAMA guarantees convergence to a balanced trade-off among conflicting objectives, offering a provably convergent and computationally efficient approach to multi-objective alignment for LLMs.

3 Experiments

In this section, we aim to empirically validate whether the theoretical advantages of PAMA are reflected in practical experiments. To this end, we conduct systematic evaluations across different model scales and diverse, potentially conflicting objectives to assess PAMA's effectiveness in multi-objective alignment.

We conduct experiments on three progressively larger language models: GPT-2 (125M), GPT-2 XL (1.5B), and LLaMA-2 (7B), and evaluate PAMA using a range of reward models, including harmlessness, humor, sentiment, and response length. Our implementation is based on the open-source TRL framework [29]. All experiments are conducted on a workstation equipped with an Intel i9-14900K CPU and a single NVIDIA RTX A6000 GPU. Further experimental details are provided in Appendix H, with additional results included in Appendix I.



(a) Length (b) Sentiment

Fig. 1: Comparison of sentiment and length rewards during training on the IMDb dataset using GPT-2 (125M parameters). PAMA consistently achieves superior performance across both objectives, demonstrating stable optimization. In contrast, MORLHF struggles to balance sentiment and length due to the limitations of the fixed weighted sum approach, while MGDA-UB does not show any advantage over MORLHF. The shaded area represents the standard deviation over eight trials, highlighting the robustness of PAMA.

3.1 Normal Model: GPT-2 (125M Parameters)

In this experiment, we evaluate PAMA on a normal-scale language model, GPT-2 (125M parameters), to assess its effectiveness in optimizing multiple objectives. Specifically, we aim to generate film reviews that are both positive and long, requiring the model to balance sentiment and length objectives.

Setup. We use GPT-2 [20] as the base model and train it on the IMDb dataset[1]. The objective consists of two reward functions: i) a pretrained sentiment analysis model[2], where the logit output serves as the reward signal to encourage positive reviews, and ii) a length-based reward that promotes longer responses. Both reward values are structured such that higher scores indicate better performance.

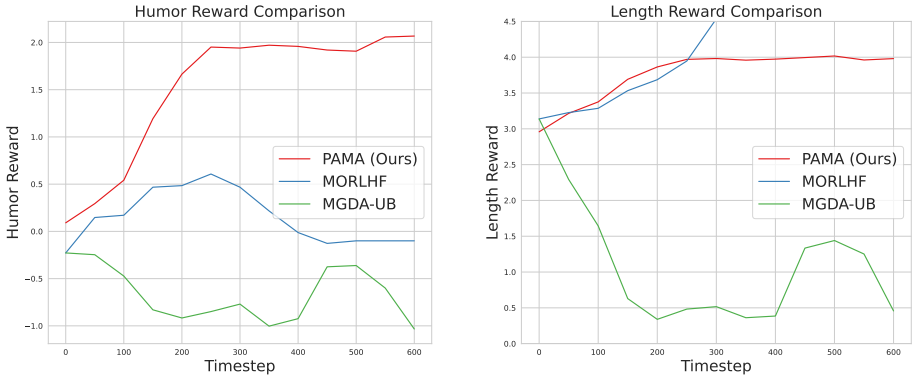
Baselines. We compare PAMA against two widely used baselines: MORLHF, which applies a fixed weighted sum of the objectives, a common but often suboptimal approach for balancing conflicting goals; and MGDA-UB [24], which leverages the min-norm algorithm to compute gradients that balance multiple objectives dynamically. Further discussion is provided in Appendix F.

Results. The training curves in Figure 1 illustrate the performance of different methods over time. Figure 1a shows that PAMA significantly outperforms both baselines in optimizing the length reward. While MORLHF and MGDA-UB

[1] https://huggingface.co/datasets/stanfordnlp/imdb
[2] https://huggingface.co/lvwerra/distilbert-imdb

exhibit slow and marginal improvements, PAMA achieves a much higher final reward with a stable convergence pattern. Figure 1b further highlights PAMA's advantage in optimizing sentiment, where it reaches a substantially higher reward than the baselines. In contrast, MORLHF stagnates at a lower level, and MGDA-UB shows negative improvement over MORLHF.

3.2 Scaling Up: GPT-2 XL (1.5B Parameters)



(a) Humor (b) Length

Fig. 2: Comparison of humor and length rewards during training on the HH-RLHF dataset using GPT-2 XL (1.5B parameters). PAMA consistently outperforms the baselines in both objectives, demonstrating stable optimization. While MORLHF fails to significantly improve humor. MGDA-UB struggles in both objectives, showing severe performance degradation. These results highlight the effectiveness of PAMA in multi-objective alignment for LLMs.

To evaluate PAMA's scalability and adaptability, we extend our experiments to GPT-2 XL (1.5B parameters), optimizing for both humor and text length.

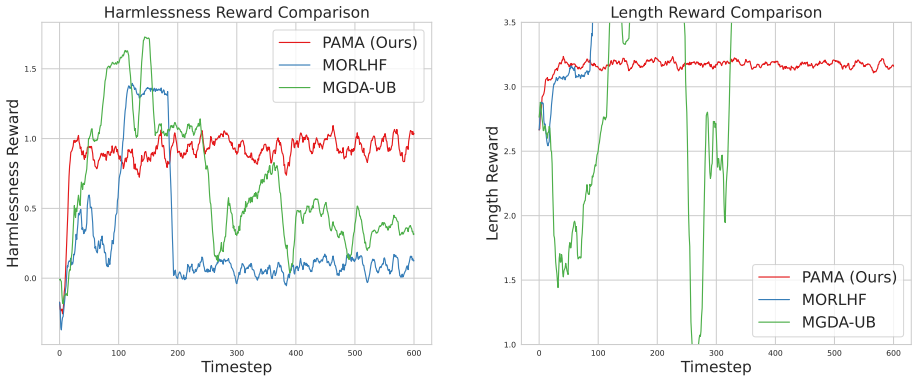
We train GPT-2 XL on the HH-RLHF dataset [2] while optimizing two distinct reward signals: i) a humor classifier[3], which assigns higher rewards to funnier outputs, and ii) a length-based reward that promotes longer responses. Higher reward values correspond to better performance in both objectives. We compare PAMA against MORLHF and MGDA-UB.

Results. The evaluation results, shown in Figure 2, illustrate the performance on the test set for humor and length rewards over training timesteps. Figure 2a demonstrates that PAMA effectively optimizes humor, steadily increasing its reward and maintaining a high final value. In contrast, MORLHF shows only marginal improvement before plateauing at a lower level, while MGDA-UB fails

[3] https://huggingface.co/mohameddhiab/humor-no-humor

entirely, with its humor reward even decreasing over time. Figure 2b shows that both PAMA and MORLHF successfully optimize length, though MORLHF only optimizes length, ignoring humor. MGDA-UB, on the other hand, completely collapses in this setting, with its length reward deteriorating throughout training. These findings reinforce PAMA's robustness in multi-objective alignment, particularly in balancing competing rewards while ensuring stable convergence.

3.3 Towards Large Language Models: LLaMA-2 7B



(a) Harmlessness (b) Length

Fig. 3: Comparison of harmlessness and length rewards during training on the HH-RLHF dataset using LLaMA-2 (7B parameters). PAMA consistently optimizes both objectives while maintaining a stable learning process. In contrast, MGDA-UB and MORLHF struggle with harmlessness optimization, exhibiting significant fluctuations and instability. MGDA-UB, in particular, exhibits pronounced oscillations during training. While MORLHF converges to a lower performance level. These results highlight the robustness of PAMA in aligning large-scale LLMs with multiple objectives.

To assess the scalability of PAMA, we extend our evaluation to a large language model setting using LLaMA-2 [26] with 7B parameters. This experiment focuses on aligning the model to generate responses that are both harmless and as long as possible. We utilize the HH-RLHF dataset and measure harmlessness using an open-source reward model[4].

Results. The results in Figure 3 demonstrate PAMA's effectiveness in large-scale multi-objective alignment. As shown in Figure 3a, PAMA achieves a stable increase in harmlessness reward, while MORLHF and MGDA-UB suffer from

[4] https://huggingface.co/Ray2333/gpt2-large-harmless-reward_model

instability and fluctuations. MGDA-UB, in particular, exhibits pronounced training oscillations, failing to maintain a high harmlessness score, whereas MORLHF stabilizes at a lower reward level. Similarly, Figure 3b illustrates that PAMA maintains strong performance in length optimization, achieving stable convergence. In contrast, MGDA-UB experiences erratic fluctuations, and MORLHF fails to sustain meaningful progress. These findings reinforce PAMA's theoretical advantages, demonstrating its ability to effectively balance competing objectives in large-scale LLM alignment.

3.4 Discussion

Our experimental results confirm that the theoretical advantages of PAMA are consistently realized in practice. Across various model size (ranging from 125M to 7B) and objective settings, PAMA demonstrates superior stability and optimization performance, significantly outperforming existing baseline methods. MORLHF, which relies on a weighted sum of objectives, struggles to balance competing rewards due to its fixed weight assignments, often leading to suboptimal trade-offs. MGDA-UB, while employing dynamic gradient balancing, can exhibit training instability and, in some cases, underperform compared to MORLHF. These findings highlight PAMA's robustness in achieving stable and well-balanced optimization across different model scales and reward settings, making it a reliable and scalable solution for multi-objective alignment in large language models.

4 Related Work

Multi-Objective Optimization is a fundamental problem in RL and deep learning, where multiple conflicting objectives must be simultaneously optimized, because improving one often leads to the degradation of another. Classical MOO techniques aim to find Pareto-optimal solutions. Among them, simple linearization methods with fixed weights often fail to effectively balance competing objectives [3]. A more general approach is Pareto-based optimization, which seeks to optimize all objectives simultaneously while maintaining trade-offs. Gradient-based MOO methods, e.g. MGDA [4], formulate a common descent direction for all objectives, ensuring simultaneous progress. However, despite their theoretical appeal, these approaches, along with related methods like PCGrad [32] and CAGrad [14], suffer from computational inefficiencies in high-dimensional parameter spaces, particularly in deep learning. The prohibitive cost of computing and aggregating gradients at LLM scale motivate the development of scalable alternatives, such as our proposed method, PAMA.

MORL extends RL to settings where an agent must learn policies that balance multiple reward functions. Standard MORL approaches include linear scalarization [27], Envelope Q-Learning [31], and Pareto Q-learning [17], as well as several recent extensions [1, 8, 10, 12, 15, 22]. These methods are widely used in applications that require trade-offs between competing objectives [7]. However, their extension to large-scale neural networks, particularly LLMs, remains an

open challenge due to computational constraints and the difficulty of balancing conflicting reward signals. A further discussion is provided in Appendices F and G.

MOO for LLMs. Applying MOO to LLMs presents additional challenges due to their high-dimensional parameter space and the inherent conflicts between objectives such as fluency, factual accuracy, and safety. Existing MOO techniques often become impractical for LLMs due to the prohibitive cost of computing gradients for each objective. For example, MGDA-UB [24] is proposed as an efficient approximation method, though its behavior on large-scale models can be unstable in practice, as observed in our experiments. Independent Component Alignment (ICA) [25] has been explored in multi-task learning for vision models, but its reliance on singular value decomposition introduces numerical instability, particularly when applied to `float16` or `bfloat16` formats used in LLM training. A notable recent approach is MOC [11], which trains an LLM as a meta-policy to generate responses aligned with user-defined preferences along the Pareto front. While promising, such approaches still face scalability and optimization challenges when applied to billion-parameter models.

Our approach, PAMA, distinguishes itself from previous methods by: i) Achieving computational efficiency comparable to single-objective RLHF methods, making it scalable to large models. ii) Providing theoretical guarantees of convergence, ensuring stable and reliable optimization. iii) Directly enabling multi-objective alignment in LLMs without relying on computationally expensive gradient manipulation techniques. By addressing both theoretical and practical limitations of existing methods, PAMA establishes a scalable and principled solution for aligning LLMs with multiple human values.

5 Conclusion

In this paper, we introduced Pareto Multi-Objective Alignment, a computationally efficient and theoretically grounded algorithm designed to align large language models across multiple, potentially conflicting objectives. By transforming the inherently complex multi-objective reinforcement learning from human feedback problem into a convex optimization framework, PAMA significantly reduces computational complexity, from an impractical $\mathcal{O}(n^2d)$ to $\mathcal{O}(n)$, where d is the number of parameters (billions for LLMs) and n is the number of objectives. This efficiency enables practical multi-objective optimization even for billion-parameter models, expanding the applicability of LLMs across diverse real-world tasks. From a theoretical perspective, we provided rigorous proofs demonstrating that PAMA converges to Pareto stationary points. The empirical results further substantiate that PAMA not only exhibits theoretical superiority but also achieves stable and efficient multi-objective alignment in real-world applications. By successfully translating its methodological advantages into tangible performance improvements, PAMA provides a computationally efficient and theoretically grounded solution for multi-objective alignment for LLMs. In summary, PAMA not only addresses a critical gap in current multi-objective alignment

methodologies but also offers a scalable, principled, and computationally viable solution for aligning LLMs with multiple human values. By establishing a strong foundation for efficient multi-objective optimization, PAMA paves the way for more adaptable, responsive, and socially aligned AI systems.

Acknowledgments. This research was supported by the German Federal Ministry of Research, Technology and Space under Grant Number 16KISK035.

Bibliography

[1] Alegre, L.N., Bazzan, A.L., Roijers, D.M., Nowé, A., da Silva, B.C.: Sample-efficient multi-objective learning via generalized policy improvement prioritization. arXiv preprint arXiv:2301.07784 (2023)

[2] Bai, Y., Jones, A., Ndousse, K., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. CoRR **abs/2204.05862** (2022), https://doi.org/10.48550/ARXIV.2204.05862

[3] Boyd, S.P., Vandenberghe, L.: Convex optimization. Cambridge university press (2004)

[4] Désidéri, J.A.: Multiple-Gradient Descent Algorithm (MGDA). Research Report RR-6953 (Jun 2009)

[5] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL-HLT 2019, pp. 4171–4186, Association for Computational Linguistics (2019)

[6] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

[7] Felten, F., Alegre, L.N., et al.: A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In: Proceedings of the 37th Conference on Neural Information Processing Systems (2023)

[8] Felten, F., Talbi, E., Danoy, G.: Multi-objective reinforcement learning based on decomposition: A taxonomy and framework. J. Artif. Intell. Res. **79**, 679–723 (2024), https://doi.org/10.1613/JAIR.1.15702

[9] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

[10] He, Q., Su, H., Zhang, J., Hou, X.: Frustratingly easy regularization on representation can boost deep reinforcement learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 20215–20225, IEEE (2023), https://doi.org/10.1109/CVPR52729.2023.01936, URL https://doi.org/10.1109/CVPR52729.2023.01936

[11] He, Q., Yang, Y., Zhou, T., Fang, M., Maghsudi, S.: One model for all: Multi-objective controllable language models (2025)

[12] He, Q., Zhou, T., Fang, M., Maghsudi, S.: Adaptive regularization of representation rank as an implicit constraint of bellman equation. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, OpenReview.net (2024), URL https://openreview.net/forum?id=apXtolxDaJ

[13] Ilyas, A., Engstrom, L., et al.: A closer look at deep policy gradients. In: 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net (2020)

[14] Liu, B., Liu, X., Jin, X., Stone, P., Liu, Q.: Conflict-averse gradient descent for multi-task learning. In: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021,, pp. 18878–18890 (2021)
[15] Lu, H., Herman, D., Yu, Y.: Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net (2023)
[16] Maas, A.L., Daly, R.E., et al.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150, Association for Computational Linguistics, Portland, Oregon, USA (June 2011)
[17] Moffaert, K.V., Nowé, A.: Multi-objective reinforcement learning using sets of pareto dominating policies. J. Mach. Learn. Res. **15**(1), 3483–3512 (2014)
[18] OpenAI: GPT-4 technical report. CoRR **abs/2303.08774** (2023), https://doi.org/10.48550/ARXIV.2303.08774
[19] Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. In: Annual Conference on Neural Information Processing Systems 2022, New Orleans, LA, USA (2022)
[20] Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
[21] Ramé, A., Couairon, G., Dancette, C., et al.: Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In: Annual Conference on Neural Information Processing Systems 2023, New Orleans, LA, USA (2023)
[22] Reymond, M., Bargiacchi, E., Nowé, A.: Pareto conditioned networks. In: 21st International Conference on Autonomous Agents and Multiagent Systems, Auckland, New Zealand, May 9-13, 2022, pp. 1110–1118 (2022)
[23] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. CoRR **abs/1707.06347** (2017)
[24] Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada, pp. 525–536 (2018)
[25] Senushkin, D., Patakin, N., Kuznetsov, A., Konushin, A.: Independent component alignment for multi-task learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, pp. 20083–20093, IEEE (2023)
[26] Touvron, H., Martin, L., Stone, K., Albert, P., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
[27] Van Moffaert, K., Drugan, M.M., Nowé, A.: Scalarized multi-objective reinforcement learning: Novel design techniques. In: 2013 IEEE Symposium on ADPRL, pp. 191–199 (2013)
[28] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Annual Conference on Neural Information Processing Systems 2017 ,Long Beach, CA, USA, pp. 5998–6008 (2017)
[29] von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S.: Trl: Transformer reinforcement learning. https://github.com/huggingface/trl (2020)

[30] Yang, R., Pan, X., Luo, F., et al.: Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In: Forty-first International Conference on Machine Learning, Vienna, Austria, July 21-27, 2024, OpenReview.net (2024)
[31] Yang, R., Sun, X., Narasimhan, K.: A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, pp. 14610–14621 (2019)
[32] Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. In: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, virtual (2020)