THE UNIVERSITY OF CHICAGO

# TOWARDS ROBUST ALIGNMENT OF LANGUAGE MODELS WITH HUMAN PREFERENCES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
CHAOQI WANG

CHICAGO, ILLINOIS
MARCH 2025

# ABSTRACT

The rapid advancement of large language models (LLMs) offers transformative potential across a wide range of applications, but concurrently raises critical safety concerns, underscoring the importance of aligning AI systems with human values. This thesis investigates a series of methodologies designed to enhance AI alignment through novel optimization strategies that improve performance, robustness, and trustworthiness. By addressing limitations in traditional reinforcement learning from human feedback (RLHF) pipelines, our work introduces streamlined alternatives and complementary techniques to optimize alignment while maintaining computational efficiency and effectiveness.

In **Chapter 2**, we propose *f-DPO* (generalized Direct Preference Optimization), which extends the DPO framework by incorporating diverse divergence constraints, such as Jensen-Shannon and forward KL divergences. Through an analytical exploration of the Karush-Kuhn-Tucker conditions, we derive simplified relationships between reward functions and optimal policies under these divergences. Empirical evaluations demonstrate that f-DPO balances alignment performance and generative diversity more effectively than RLHF-based Proximal Policy Optimization (PPO). It also achieves lower expected calibration error (ECE) while providing practical benefits, such as improved divergence efficiency.

**Chapter 3** extends alignment optimization by addressing the limitations of single-sample preference comparison. We introduce *Multi-sample Direct Preference Optimization (mDPO)* and *Multi-sample Identity Preference Optimization (mIPO)*, frameworks that optimize group-wise characteristics to improve distributional properties such as diversity and bias reduction. These approaches significantly enhance generative diversity in LLMs and mitigate demographic biases in diffusion models. Moreover, multi-sample methods exhibit robustness to noisy human-labeled preference data, making them particularly effective for fine-tuning in real-world scenarios where label quality may be imperfect. More importantly, it offers more controllability for improving the alignment of generative models over f-dpo.

In **Chapter 4**, we move beyond the traditional supervised learning in chapters 2 and 3, and address a critical challenge in reward modeling for online RLHF: spurious correlations that can distort alignment objectives. We introduce a *causal reward modeling* framework that integrates causal inference techniques to mitigate these biases. By enforcing counterfactual invariance, this approach ensures reward predictions remain unaffected by irrelevant or confounding variables. Experiments on synthetic and real-world datasets demonstrate significant improvements in addressing biases such as length preference, sycophancy, and concept biases, ultimately enhancing the fairness and reliability of alignment.

Together, these contributions advance the theoretical and practical foundations of AI alignment. By offering scalable and robust methodologies, this thesis bridges the gap between current capabilities and the long-term goal of developing safe and trustworthy AI systems. The proposed approaches not only improve alignment workflows but also address critical shortcomings in existing pipelines, paving the way for more reliable, diverse, and human-aligned AI systems.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

The journey of a PhD is a unique blend of excitement, anxiety, and uncertainty. It is a time of intense self-motivation—whether brainstorming ideas, learning new skills, writing papers, coding, or conducting experiments. Looking back, this experience has been deeply rewarding, offering opportunities to think critically, challenge myself, and explore topics that spark my curiosity.

This journey would not have been possible without the support of many people. First and foremost, I would like to express my heartfelt gratitude to my advisor, Prof. Yuxin Chen. Yuxin has been an incredible mentor throughout my PhD. His unwavering support and the freedom he provided to explore my research interests were invaluable. He always welcomed discussions and offered thoughtful guidance with kindness and patience. Yuxin's trust in my independence has given me the confidence to explore my ideas and push my boundaries, and for that, I am profoundly grateful.

I am also deeply grateful to my thesis committee members, Prof. Haifeng Xu and Prof. Ari Holtzman, for their insightful and constructive feedback, which has been instrumental in shaping this thesis. My collaboration with Haifeng on online learning, which began as a course project, evolved into a rewarding opportunity to contribute to a spotlight at NeurIPS. Through this, I learned valuable techniques for proving regret bounds and working with concentration inequalities, all thanks to Haifeng's profound insights. His depth of understanding and sharp perspective have left a lasting impression on me, and I am truly grateful for the opportunity to learn from him.

I would also like to thank Drs. Kevin Murphy and Sinong Wang for offering me the chance to intern at Google Brain and Meta GenAI. My first exposure to JAX and model-based reinforcement learning during these internships was invaluable. Kevin's support and dedication to science, especially Bayesian methods, were truly inspiring. His passion for research and his deep understanding of the field have been a guiding light for my own work. I

also owe a debt of gratitude to Sinong for his mentorship at Meta. As my internship manager, Sinong created an environment of support and inspiration, and working on language models with him was not only intellectually rewarding but also incredibly enjoyable.

Along the way, I have been fortunate to forge lasting friendships with my labmates and colleagues, who made my PhD journey both enriching and enjoyable. Yibo Jiang, Zhuokai Zhao, Zhaorun Chen, Ziyu Ye, and Chenghao Yang and many others have been essential to making this experience memorable. From collaborating on research projects to sharing lighthearted moments during stressful times, their companionship made all the difference. I could not have asked for a better group of people to share this journey with, and I am deeply thankful for each of them. The camaraderie and support we have given each other have made this PhD not only bearable but truly rewarding.

I am also incredibly grateful for the friendships I formed during my time in Toronto with Guodong Zhang, Shengyang Sun, and Ricky Chen. Although we were at different locations, we stayed connected through our shared experiences and discussion. Our continued discussions about research, ideas, and challenges have been invaluable in shaping my own approach to academic work. From them, I have learned not only the technical aspects of conducting research but also the importance of setting high standards, questioning assumptions, and striving for excellence in all things. Our friendship has been a source of motivation and a reminder that great research is built on the foundation of meaningful collaboration and shared knowledge.

Lastly, I would like to express my deepest gratitude to my family, whose support has been unwavering throughout this entire journey. To my parents, thank you for your boundless love and encouragement, which have kept me going through the most difficult times. Your belief in me has been a constant source of strength, and your sacrifices have made this journey possible. I cannot thank you enough for being my pillar of support, and I dedicate this accomplishment to you.

# CHAPTER 1

# INTRODUCTION

Artificial intelligence (AI) has emerged as a transformative force reshaping industries, research, and daily life. From automating routine tasks to solving complex, previously intractable problems, AI continues to redefine the boundaries of technological possibility [OpenAI, 2023b, Touvron et al., 2023b, Gemini et al., 2023]. Central to this revolution is the development of machine learning (ML) systems, particularly deep learning (DL) models, which enable AI to learn from vast amounts of data, uncover intricate patterns, and make predictions or decisions without relying on explicitly programmed instructions [Christiano et al., 2017, Bai et al., 2022b, Touvron et al., 2023b]. The remarkable progress in DL has been instrumental in unlocking groundbreaking advancements across diverse fields such as computer vision, natural language processing, recommender systems, and more.

Despite these successes, the rapidly increasing capabilities of AI systems have also introduced significant challenges, particularly in ensuring that such systems are aligned with human values and operate safely. The field of AI alignment has gained prominence as researchers and practitioners seek to address risks associated with misaligned AI, including potential misuse in areas such as social manipulation [Hendrycks et al., 2023], cybersecurity threats [Shevlane et al., 2023], and biotechnological hazards [Urbina et al., 2022]. These risks underscore the urgency of developing robust methodologies to align AI systems with human intentions and ensure their responsible deployment.

Reinforcement Learning from Human Feedback (RLHF) [Christiano et al., 2017] has emerged as a promising framework for addressing these alignment challenges. By incorporating human preferences into the training pipeline, RLHF enables AI systems to exhibit behavior consistent with human expectations and values. However, the RLHF paradigm is not without its limitations. The reliance on reward models, which infer human preferences, introduces vulnerabilities to spurious correlations and reward hacking, while reinforcement

learning algorithms such as Proximal Policy Optimization (PPO) [Schulman et al., 2017] present stability and efficiency concerns.

This thesis explores alternatives and extensions to the RLHF framework, focusing on addressing its inherent challenges and broadening its applicability to diverse alignment scenarios. Central to the proposed approaches is the development of novel algorithms and methodologies that optimize data utilization, enhance generative diversity, and improve the robustness of alignment processes. By integrating advanced techniques such as divergence constraints, multi-sample comparisons, and robust evaluation pipelines, this work aims to contribute to the next generation of AI systems that are not only powerful but also trustworthy and aligned with human values.

### *1.0.1 Thesis Structure*

The primary objective of this thesis is to advance AI alignment through innovative approaches that overcome the limitations of existing methodologies. Specifically, this work focuses on:

- **Chapter 2**: focuses on divergence constraints in RLHF, introducing a generalized framework that enables more flexible alignment strategies by leveraging diverse divergence measures.
- **Chapter 3**: Explores multi-sample comparison techniques to enhance post-training processes, ensuring better control over generative diversity and mitigating biases in model outputs.
- **Chapter 4**: Investigates the challenges of reward modeling, proposing methods to identify and correct spurious correlations and reward hacking, thereby improving alignment robustness.

By addressing critical challenges in AI alignment, this thesis contributes to the broader

effort of developing safe, reliable, and human-aligned AI systems. The methodologies and findings presented herein have implications for improving the robustness, generalizability, and ethical deployment of AI across various domains. Moreover, this work provides a foundation for future research into scalable and trustworthy AI alignment strategies, ensuring that AI technologies serve as a force for societal good.

### *1.0.2 Publications Relevant to This Thesis*

The following publications [Wang et al., 2023a, 2024c, 2025] directly contribute to Chapters 2, 3, and 4 of this thesis, forming its main body.

- **Chaoqi Wang**, Yibo Jiang, Chenghao Yang, Han Liu, Yuxin Chen. Beyond Reverse KL: Generalizing Direct Preference Optimization with Diverse Divergence Constraints. *In ICLR 2024, Spotlight.* https://arxiv.org/pdf/2309.16240

- ***Chaoqi Wang**, Zhuokai Zhao, Chen Zhu, Karthik Abinav Sankararaman, Michal Valko, Xuefei Cao, Zhaorun Chen, Madian Khabsa, Yuxin Chen, Hao Ma, Sinong Wang. Preference optimization with multi-sample comparisons. In arXiv preprint arXiv:2410.12138.* https://arxiv.org/pdf/2410.12138

- ***Chaoqi Wang**, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Xiangjun Fan, Hao Ma, Sinong Wang. Beyond Spurious Correlations: Causal Rewards for Language Model Alignment.* https://arxiv.org/pdf/2501.09620

Additionally, the content from the following publications [Wang et al., 2024a,b, 2023b, 2021a] is relevant to this thesis, particularly the sections on reinforcement learning and online learning. However, their focus is not on language models but rather on traditional tasks such as robotics and online recommendation systems. As a result, they are not directly included in this thesis.

- **Chaoqi Wang**, Yuxin Chen, Kevin Murphy. Model-based Policy Optimization under Approximate Bayesian Inference. *In AISTATS, 2024, Oral.* https://proceedings.mlr.press/v238/wang24g/wang24g.pdf

- ***Chaoqi Wang***, *Ziyu Ye, Kevin Murphy, Yuxin Chen. Don't Be Pessimistic Too Early: Look K Steps Ahead! In AISTATS, 2024* https://proceedings.mlr.press/v238/wang24h/wang24h.pdf

- ***Chaoqi Wang***, *Ziyu Ye, Zhe Feng, Ashwinkumar Badanidiyuru, Haifeng Xu. Follow-ups Also Matter: Improving Contextual Bandits via Post-serving Contexts. In NeurIPS, 2023, Spotlight* https://arxiv.org/pdf/2309.13896

- ***Chaoqi Wang***, *Adish Singla, Yuxin Chen. Teaching an Active Learner with Contrastive Examples. In NeurIPS, 2021.* https://arxiv.org/pdf/2110.14888

# CHAPTER 2

# GENERALIZING DIRECT PREFERENCE OPTIMIZATION WITH DIVERSE DIVERGENCE CONSTRAINTS

## 2.1 Introduction

The increasing capabilities of large language models (LLMs) raise opportunities for artificial general intelligence but concurrently amplify safety concerns, such as potential misuse of AI systems, necessitating effective AI alignment. Reinforcement Learning from Human Feedback (RLHF) has emerged as a promising pathway towards AI alignment but brings forth challenges due to its complexity and dependence on a separate reward model. Direct Preference Optimization (DPO) has been proposed as an alternative, and it remains equivalent to RLHF under the reverse KL regularization constraint. This chapter presents $f$-DPO, a generalized approach to DPO by incorporating diverse divergence constraints. We show that under certain $f$-divergences, including Jensen-Shannon divergence, forward KL divergences and $\alpha$-divergences, the complex relationship between the reward and optimal policy can also be simplified by addressing the Karush–Kuhn–Tucker conditions. This eliminates the need for estimating the normalizing constant in the Bradley-Terry model and enables a tractable mapping between the reward function and the optimal policy. Our approach optimizes LLMs to align with human preferences in a more efficient and supervised manner under a broad set of divergence constraints. Empirically, adopting these divergences ensures a balance between alignment performance and generation diversity. Importantly, $f$-DPO outperforms PPO-based methods in divergence efficiency, and divergence constraints directly influence expected calibration error (ECE).

## 2.2 Divergence Constraints in RLHF

The increasing capabilities of large language models [Bubeck et al., 2023, OpenAI, 2023a] hold promise for achieving artificial general intelligence. However, they also pose safety concerns within the scope of AI risk [Amodei et al., 2016, Hendrycks et al., 2023]. Some of the hazardous capabilities an AI system may possess include social manipulation [Hendrycks et al., 2023], AI-enabled cyberattacks [Shevlane et al., 2023], and enhanced pathogens [Urbina et al., 2022]. These could be misused by humans or exploited by the AI system itself. Consequently, AI alignment research becomes critically important in ensuring AI systems are robustly aligned with human values.

Reinforcement Learning from Human Feedback (RLHF) has emerged as a concrete research agenda, proving effective in aligning model behaviors with human preferences and instruction following [Christiano et al., 2017, Bai et al., 2022a, Touvron et al., 2023a]. Given the challenge of specifying an objective that accurately represents human preferences in RLHF, researchers typically collect a dataset that reflects human preference in terms of model-wise generation comparisons [Bai et al., 2022a, LAION-AI, 2023]. Subsequently, a reward model is trained under the Bradley-Terry model [Bradley and Terry, 1952] to infer the human's objective from the collected dataset. The language model is then fine-tuned using RL algorithms such as Proximal Policy Optimization [Schulman et al., 2017, Ouyang et al., 2022] or Advantage Actor-Critic [Mnih et al., 2016, Glaese et al., 2022] to maximize the reward. This process is carried out while ensuring that the model does not deviate significantly from its original form, using the reverse KL divergence penalty.

While effective, the RLHF pipeline is significantly more complex than supervised learning. In particular, RLHF necessitates training a separate reward model. The quality of this model ultimately determines the performance of reinforcement fine-tuning, and the language model may exploit errors present within the reward model [Gao et al., 2023]. Additionally, RL algorithms, such as PPO [Schulman et al., 2017], are less stable and more memory-demanding

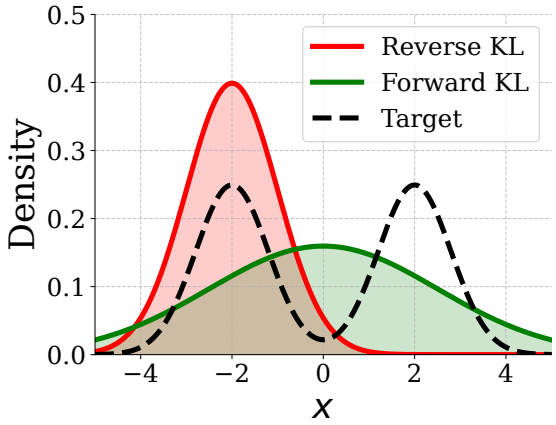


Figure 2.1: The mode seeking and mass covering behaviors of reverse KL and forward KL.

than supervised learning [Touvron et al., 2023a]. These challenges pique interest in searching for alternatives to the RLHF pipeline. Such efforts include Reward rAnked FineTuning (RAFT) [Dong et al., 2023], Rank Responses to align Human Feedback (RRHF) [Yuan et al., 2023], and Direct Preference Optimization (DPO) [Rafailov et al., 2023]. DPO, as an early initiative, leverages the mapping between the reward function and the optimal policy to bypass the need for reinforcement learning and explicit reward model learning. Still, it's equivalent to RLHF for the final solution under reverse KL regularization.

However, most existing studies focus on solutions under the constraint of reverse KL divergence, and the exploration of incorporating other divergences remains significantly lacking. To illustrate the differences among various divergence constraints, we have visualized the mode-seeking and mass-covering behaviors of reverse KL and forward KL divergences in Figure 2.1

The mode-seeking property of reverse KL divergence tends to reduce diversity in generation [Wiher et al., 2022, Khalifa et al., 2021, Perez et al., 2022b, Glaese et al., 2022], which, although beneficial for optimizing alignment performance, may limit the model's potential (e.g., user engagement). Additionally, Santurkar et al. [2023] observe that finetuning LLMs with RLHF under the reverse KL regularization will result in a limited range of political views. As a mitigation, the inclusion of various divergences can lead to solutions with

distinct characteristics (e.g., diversity), especially when the model is under-specified.

In this chapter, we generalize the DPO framework [Rafailov et al., 2023] to incorporate various divergence constraints. Different from the reverse KL divergence, we initially find that a naive derivation results in an excessively complex relationship between the reward and optimal policy for other divergences, largely due to the normalizing constant. However, by meticulously addressing the Karush–Kuhn–Tucker (KKT) conditions, specifically the complementary slackness, we demonstrate that for a class of well-known divergences, such as Jensen-Shannon divergences, forward KL divergences and $\alpha$-divergences with $\alpha \in (0, 1)$, the normalizing constant can be eliminated in the Bradley-Terry model. This results in an analytical and elegant mapping between the reward function and the optimal policy for a broad class of divergences. It further enables us to optimize the language model to align with human preferences under varying divergences constraints without needing to estimate the normalization constant. Such flexibility might allow us to explore a richer spectrum of modeling possibilities and cater to diverse application requirements

In conclusion, our key contribution is the generalization of the DPO framework to seamlessly integrate a variety of popular divergences (e.g., Jensen-Shannon divergence, forward KL divergence and $\alpha$-divergence) for regularization. Empirical results indicate that by adapting different divergence regularizations, we can achieve a nuanced balance between alignment performance (e.g., reward) and the generation diversity. This introduces greater flexibility during fine-tuning processes utilizing human preference datasets. Furthermore, comparative analyses reveal that the generalized DPO framework surpasses PPO-based methods in divergence efficiency. We also prove that the difference in the expected calibration error (ECE) is bounded by the divergence between them, emphasizing the practical benefits of improved divergence efficiency in model calibration.

## 2.3 Related Works

**AI Alignment** [Leike et al., 2018] is proposed as a research agenda aimed at aligning model behavior with human preferences and instruction following. Not only has AI alignment been demonstrated to be essential in ensuring safe AI behaviors, but it also improves performance on a variety of downstream tasks [Bai et al., 2022a,b, OpenAI, 2023a, Touvron et al., 2023a, Glaese et al., 2022]. These tasks include metrics such as helpfulness [Askell et al., 2021], truthfulness [Lin et al., 2022], and non-offensiveness [Gehman et al., 2020], etc. In this context, numerous methodologies have been proposed, including red teaming [Perez et al., 2022b, Korbak et al., 2023], reward modeling [Leike et al., 2018, Gao et al., 2023], supervised fine-tuning, rejection sampling, and reinforcement learning from human/AI feedback [Ziegler et al., 2019, Ouyang et al., 2022, Bai et al., 2022b], among others. These methods largely depend on human judgment or a comprehensive set of human-written principles to provide the supervised signal. For more complex situations where humans may be incapable of evaluating, the main approach involves designing mechanisms that utilize AI to assist in evaluation by recursively decomposing the problem. This body of work includes AI debate [Irving et al., 2018], iterated amplification [Christiano et al., 2018], and recursive reward modeling [Leike et al., 2018]. However, most of these methods typically involve multiple stages of training or complex interaction protocols. In contrast, we focus on proposing a single-stage algorithm that is simple to implement and computationally efficient in the setup where a human is capable of evaluating.

**Reinforcement Learning from Human Feedback (RLHF)** [Christiano et al., 2017, Bai et al., 2022a, Touvron et al., 2023a, Ouyang et al., 2022] has served as a pivotal method for aligning language models, contributing significantly to the success of ChatGPT [OpenAI, 2023a]. Nonetheless, RLHF's complexity surpasses that of supervised learning, primarily due to the need for a distinct reward model, the quality of which decisively influences the efficacy of reinforcement fine-tuning. Any errors within this model may be exploited by

the language model [Gao et al., 2023]. Furthermore, RL algorithms like PPO [Schulman et al., 2017] have been proven to be less stable and more memory-intensive than supervised learning [Touvron et al., 2023a]. These challenges have spurred interest in alternatives to the RLHF pipeline, such as Reward Ranked FineTuning (RAFT) [Dong et al., 2023], Rank Responses to align Human Feedback (RRHF) [Yuan et al., 2023], and Direct Preference Optimization (DPO) [Rafailov et al., 2023]. Despite these efforts, all the methods mentioned focus solely on solutions within the confines of reverse KL divergence regularization, leaving the potential advantages of incorporating various other divergences largely unexplored. Go et al. [2023] recently attempted to minimize the $f$-divergence for aligning language models. However, this approach necessitates the user to define the target distribution, and requires the estimation of the target distribution's normalizing constant—adding more hyperparameters and algorithmic complexity. In contrast, we propose a supervised learning method that incorporates various divergence regularizations. This method does not require the estimation of normalizing constants, the specification of the target distribution, or the use of reinforcement learning (e.g., model rollouts). Moreover, it does not involve any additional hyperparameters.

## 2.4 Preliminary and Backgrounds

### *2.4.1 Preliminary*

**$f$-divergences.** For any convex function $f : \mathbb{R}^+ \to \mathbb{R}$ that satisfies $f(1) = 0$ and $f$ is strictly convex around 1, then the corresponding $f$-divergence for two distributions $p$ and $q$ is defined as

$$D_f(p, q) = \mathbb{E}_{q(x)} \left[ f\left( \frac{p(x)}{q(x)} \right) \right].$$

The $f$-divergence covers a broad class of commonly used divergences, including forward KL divergence, reverse KL divergence, Jensen-Shannon (JS) divergence and total variation distance, etc, by choosing the specific function $f$. We provide a summarization in Table 2.1.

**Bradley-Terry Model.** The Bradley-Terry model [Bradley and Terry, 1952] has been widely employed for pairwise comparisons. It works by assigning a real-valued "strength" parameter $p_i$ to each item $y_i$, which is then used to compute the probability that item $y_i$ outperforms item $y_j$ in a pairwise comparison by $p(y_i \succeq y_j) = p_i/(p_i + p_j)$. Yet, in practice, this model's instantiation usually adopts a specific form, denoted as

$$p(y_i \succeq y_j) = \frac{\exp(r(y_i))}{\exp(r(y_i)) + \exp(r(y_j))},$$

where $r(y_i)$ represents the "rating" or transformed "strength" of item $y_i$.

The Bradley-Terry model can be linked to Gumbel noise in the context of pairwise comparisons. The outcome depends on the comparison of $r(y_i) + \epsilon_i$ and $r(y_j) + \epsilon_j$, where $\epsilon_i$ and $\epsilon_j$ are i.i.d Gumbel-distributed noise. If $r(y_i) + \epsilon_i > r(y_j) + \epsilon_j$, $y_i$ defeats $y_j$. Noting that the difference between two Gumbel variables follows a logistic distribution, we align this with the Bradley-Terry model's formula: $p(y_i \succeq y_j) = 1/(1 + \exp((r(y_j) - r(y_i))))$, with $p_i = \exp(r(y_i))$ and $p_j = \exp(r(y_j))$. Thus, the Bradley-Terry model represents a stochastic rank-order model influenced by Gumbel noise, which may justify the use of it for categorical distribution (e.g., language models).

**RL from Human Feedbacks.** RLHF takes place after the base model has been pre-trained. It comprises three steps: 1) supervised fine-tuning, 2) reward model training, and 3) RL fine-tuning. In the RL fine-tuning process, the following objective is maximized:

$$\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{\boldsymbol{\theta}}(\cdot|x)}\left[r_{\varphi}(y|x)\right] - \beta D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}(\cdot|x)|\pi_{\mathrm{ref}}(\cdot|x)),$$

where $\mathcal{D}$ represents the dataset of prompts, $r_{\varphi}(\cdot|x)$ stands for the reward function learned

using the Bradley-Terry model on the preference dataset, $\pi_{\text{ref}}(\cdot|x)$ is the fixed reference model (typically selected to be the one post supervised fine-tuning), and $\beta$ is the coefficient of the reverse KL divergence penalty. In practice, this objective is equivalent to executing reinforcement learning under the following reward function [Ziegler et al., 2019, Bai et al., 2022a, Ouyang et al., 2022]:

$$r(\cdot|x) = r_{\varphi}(\cdot|x) - \beta \log \left( \frac{\pi_{\boldsymbol{\theta}}(\cdot|x)}{\pi_{\text{ref}}(\cdot|x)} \right).$$

**Directed Preference Optimization (DPO)**. The original DPO method [Rafailov et al., 2023] establishes a functional mapping between the reward model and the optimal policy under the reverse KL divergence constraint. This allows for the direct optimization of the policy by reparameterizing the reward function using the policy (i.e., the language model) in a supervised manner,

$$r(\cdot|x) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(\cdot|x)}{\pi_{\text{ref}}(\cdot|x)} + \beta \log Z(x).$$

Here, $Z(x)$ is the partition function or the normalizing constant. By plugging the reward into the Bradley-Terry model, the resulting objective of DPO with reverse KL divergence is given by:

$$-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_{\boldsymbol{\theta}}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} + \cancel{\beta \log Z(x)} - \beta \log \frac{\pi_{\boldsymbol{\theta}}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \cancel{\beta \log Z(x)} \right) \right],$$

where $\sigma$ is the sigmoid function, and the partition functions are cancelled out.

**Calibration Error.** To measure calibration, it's common to use the Expected Calibration Error (ECE) [Guo et al., 2017]. Specifically, for any policy $\pi_{\boldsymbol{\theta}}(\cdot|x)$, define $\hat{P}_{\pi_{\boldsymbol{\theta}}}(x)$ as

$$\hat{P}_{\pi_{\boldsymbol{\theta}}}(x) = \pi_{\boldsymbol{\theta}}(\hat{y}|x),$$

where $\hat{y}$ is the predicted label of $x$. Given a policy $\pi_{\boldsymbol{\theta}}$, $\hat{y}$ is sampled with probability $\pi_{\boldsymbol{\theta}}(\hat{y}|x)$. Then, the ECE can be computed by,

$$\mathrm{ECE}(\boldsymbol{\theta}) = \mathbb{E}_{\hat{P}_{\pi_{\boldsymbol{\theta}}}}[|\mathbb{P}(\hat{Y} = Y|\hat{P}_{\pi_{\boldsymbol{\theta}}} = p) - p|].$$

Intuitively, if the model's confidence in its predictions closely matches the probability of its predictions being correct, then the model is well-calibrated. In the realm of LLMs, OpenAI [2023a] demonstrate that RLHF degrades the model's calibration.

### *2.4.2 Background: Optimizing for Reverse KL Hurts Diversity*

The finetuning of Large Language Models (LLMs) using Reinforcement Learning from Human Feedback (RLHF) has raised concerns regarding sample diversity. Notably, this process is prone to mode collapse (see Figure 2.1 for illustration). This will result in a reduction in the diversity of model outputs, as evidenced by studies from Khalifa et al. [2021], Perez et al. [2022b], Go et al. [2023], and Glaese et al. [2022]. One plausible explanation for mode collapse is the shift from supervised to reinforcement learning with reverse KL divergence [Song et al., 2023]. Additionally, Santurkar et al. [2023] found that LLMs finetuned with RLHF under reverse KL divergence regularization can express a limited range of political views. Given these observations, there's a clear need to investigate alternative divergence regularization methods to maintain the diversity and integrity of LLM outputs, and understand the tradeoff.

## 2.5 Method: Direct Preference Optimization under $f$-divergence

During the RL finetuning process, it is common to regularize the finetuned model to stay "close" to the original model (or reference model) measured by the KL divergence. Typically, reverse KL divergence (i.e., $D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}\|\pi_{\mathrm{ref}})$) is a default choice. However, the mode-seeking

Table 2.1: Summary of some commonly used $f$-divergences including their derivatives.

| $f$-**divergence** | $\boldsymbol{f(u)}$ | $\boldsymbol{f'(u)}$ | $0 \notin$ Domain of $\boldsymbol{f'(u)}$ |
|---|---|---|---|
| $\alpha$-divergence ($\alpha \in (0,1)$) | $(u^{1-\alpha} - (1-\alpha)u - \alpha)/(\alpha(\alpha-1))$ | $(1-u^{-\alpha})/\alpha$ | ✓ |
| Reverse KL ($\alpha = 0$) | $u \log u$ | $\log u + 1$ | ✓ |
| Forward KL ($\alpha = 1$) | $-\log u$ | $-1/u$ | ✓ |
| JS-divergence | $u \log u - (u+1)\log((u+1)/2)$ | $\log(2u/(1+u))$ | ✓ |
| Total Variation | $\frac{1}{2}\lvert u-1\rvert$ | $u > 1\ ?\frac{1}{2} : -\frac{1}{2}$ | ✗ |
| Chi-squared | $(u-1)^2$ | $2(u-1)$ | ✗ |

behavior will lead to low diversity in the generations. Therefore, to balance the alignment performance (e.g., accuracy or reward) and diversity, we consider a more broad class of divergence regularization, namely, the $f$-divergence, which covers many commonly used divergences, such as forward KL, reverse KL, JS divergence and $\alpha$-divergence, etc.

When given the reward function $r(y|x)$, the base model is fine-tuned using reinforcement learning to maximize the reward under certain constraints. From an optimization perspective, this step is equivalent to solving the following constrained optimization problem:

$$\max_{\pi} \mathbb{E}_{\pi}[r(y|x)] - \beta D_f(\pi, \pi_{\text{ref}}) \quad s.t. \quad \sum_y \pi(y|x) = 1 \ \text{ and } \ \pi(y|x) \geq 0 \ \forall y.$$

The two constraints are introduced to ensure that the solution is a valid distribution, though in practice we don't need to explicitly deal with them. To solve the constrained problem, we can apply the Lagrange multiplier, which gives us

$$\mathcal{L}(\pi, \lambda, \alpha) = \mathbb{E}_{\pi}[r(y|x)] - \beta \mathbb{E}_{\pi_{\text{ref}}}\left[f\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right] - \lambda\left(\sum_y \pi(y|x) - 1\right) + \sum_y \alpha(y)\pi(y|x),$$

where $\lambda$ and $\alpha(y)$ are the dual variables. For such problems, we can derive the closed-form solution for $\pi^\star$, which optimally solves the above problem:

$$\pi^\star(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)(f')^{-1}\left(\frac{r(y|x)}{\beta}\right),$$

where $Z(x)$ is the normalization constant, and $(f')^{-1}$ is the inverse function of $f'$. By solving the equation for $r(y|x)$, we establish the following relationship between $r(y|x)$ and $\pi^\star(y|x)$,

$$r(y|x) = \beta f' \left( \frac{\pi^\star(y|x)}{\pi_{\text{ref}}(y|x)} \cdot Z(x) \right).$$

When $D_f$ is the reverse KL divergence, we have $f'(u) = \log u + 1$ from Table 2.1. Thus, $r(y|x) = \beta \log(\pi^\star(y|x)/\pi_{\text{ref}}(y|x)) + \log Z(x) + 1$. Since the BT model is defined by $p(y_w \succ y_l) = \sigma(r(y_w|x) - r(y_l|x))$, the $\log Z(x)$, which appears as an additive term in the reward, will be canceled out for the reverse KL divergence, but this is not the case for other divergences. This situation is discouraging, as estimating the $Z(x)$ requires multiple samples, which may be computationally expensive and may exhibit high variance if not handled properly.

Fortunately, we will show that, by carefully analyzing the normalization constant $Z(x)$, we can derive a closed-form solution for many other (but not all) divergences as well, without the need to estimate the normalization constant $Z(x)$. To achieve this, we can first rewrite $\pi^\star(y|x)$ in the following format using the dual variables $\lambda$ and $\alpha(y)$:

$$\pi^\star(y|x) = \pi_{\text{ref}}(y|x)(f')^{-1} \left( \frac{r(y|x) - \lambda + \alpha(y)}{\beta} \right).$$

However, the term $\alpha(y)$ depends on $y$, and cannot be canceled out, making it hard to compute. Next, we will show that for a class of $f$-divergences, we must have $\alpha(y) = 0$, and thus we can represent the reward using only the trainable policy, the reference policy, and a constant that is independent of $y$. This result is summarized in the following theorem.

**Theorem 1.** *If $\pi_{\text{ref}}(y|x) > 0$ for any valid $x$ and $f'$ is invertible with $0 \notin dom(f')$, the reward class that is consistent with the Bradley-Terry model can be reparameterized using the policy model $\pi(y|x)$ and a reference model $\pi_{\text{ref}}(y|x)$ as*

$$r(y|x) = \beta f' \left( \frac{\pi^\star(y|x)}{\pi_{\text{ref}}(y|x)} \right) + \text{const.} \tag{2.1}$$

Theorem 1 holds mainly due to the complementary slackness in the KKT conditions, and the proof can be found in Appendix 2.8.7. Moreover, the requirement $0 \notin \mathrm{dom}(f)$ already covers many commonly used divergences, including forward KL divergence, Jensen-Shannon divergence, reverse KL divergence, and $\alpha$-divergence with $0 < \alpha < 1$. Lastly, the constant is independent of $y$ and appears as an additive term in the reward function, and thus it will be cancelled out when plugged into the Bradley-Terry model.

Now, for a pair of examples $(x, y_w)$ and $(x, y_l)$, we can plug the reward from Equation 2.1 into the Bradley-Terry model, which gives us the following expression,

$$p(y_w \succeq y_l|x) = \sigma\left(\beta f'\left(\frac{\pi^\star(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)}\right) - \beta f'\left(\frac{\pi^\star(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right)\right).$$

Hence, for a preference dataset $\mathcal{D}$, we train the model $\pi_{\boldsymbol{\theta}}$ (replacing $\pi^\star$ in the above equation) by minimizing the following negative log-likelihood loss,

$$\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[-\log\sigma\left(\beta f'\left(\frac{\pi_{\boldsymbol{\theta}}(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)}\right) - \beta f'\left(\frac{\pi_{\boldsymbol{\theta}}(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right)\right)\right]. \tag{2.2}$$

The above helps us to solve the RL finetuning problem via a supervised learning approach under a broad class of divergence constraints, which is more stable and efficient to optimize in contrast to the reinforcement learning counter-part. Our results generalize the DPO [Rafailov et al., 2023] to a more broad class of divergence regularization. The full algorithm is summarized in Algorithm 1.

## 2.6 Experiments

### *2.6.1 Experimental Setup*

**Baselines and Datasets.** For the experiments, we adopt three datasets, including IMDB-sentiment dataset [Maas et al., 2011b], Anthropic HH dataset [Bai et al., 2022a] and MT-

**Algorithm 1** Direct Preference Optimization with $f$-divergences (DPO-$f$)

**Require:** Preference dataset $\mathcal{D}$, batch size $b$, constraint coefficient $\beta$, divergence function $f$, and learning rate $\eta$.

1: Initialize model $\pi_{\boldsymbol{\theta}_0}$ with supervised finetuning on $\mathcal{D}$.
2: **for** $n = 1 \dots N$ iterations **do**
3: Sample a batch $\mathcal{B} = \{(x_i, y_i^w, y_i^l)\}_{i=1}^b$ from $\mathcal{D}$.
4: Compute the loss using the equation 2.2 with the chosen function $f$.
5: Compute the gradient and update the model $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_{t-1}, \mathcal{B})$.
6: **end for**
7: **return** Final model $\pi_{\boldsymbol{\theta}}$.

bench [Zheng et al., 2023] for evaluation. Our primary baseline approach is PPO with different $f$-divergences. Upon experimentation (see Appendix 2.8.1), it was observed that incorporating a divergence penalty in the reward for variants of PPO, such as those with forward KL divergence and JS divergence, induces training instability. The reason being that the value ranges of these divergences are considerably larger than those of reverse KL divergence. This discrepancy causes substantial challenges when trying to learn a accurate value function in PPO. To address the issue, we further propose a modified variant of PPO as an additional baseline. Instead of integrating the divergence penalty into the reward function, we treat it as a regularization term separately. The new objective (i.e., PPO (loss)) is optimized separately by PPO and SGD,

$$\text{PPO (reward):} \qquad \underbrace{\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{\boldsymbol{\theta}}(\cdot|x)}\left[r_\varphi(y|x) - \beta f(\pi_{\boldsymbol{\theta}}(y|x)/\pi_{\text{ref}}(y|x))\right]}_{\text{Optimized by PPO}},$$

$$\text{PPO (loss):} \qquad \underbrace{\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{\boldsymbol{\theta}}(\cdot|x)}\left[r_\varphi(y|x)\right]}_{\text{Optimized by PPO}} - \underbrace{\beta D_f(\pi_{\text{ref}}(\cdot|x), \pi_{\boldsymbol{\theta}}(\cdot|x))}_{\text{Optimized by SGD}}.$$

For the sake of differentiation, we term the conventional PPO method as PPO (reward) – indicating the inclusion of divergence penalty in the reward. In contrast, the variant where the divergence regularization is treated separately is denoted as PPO (loss).

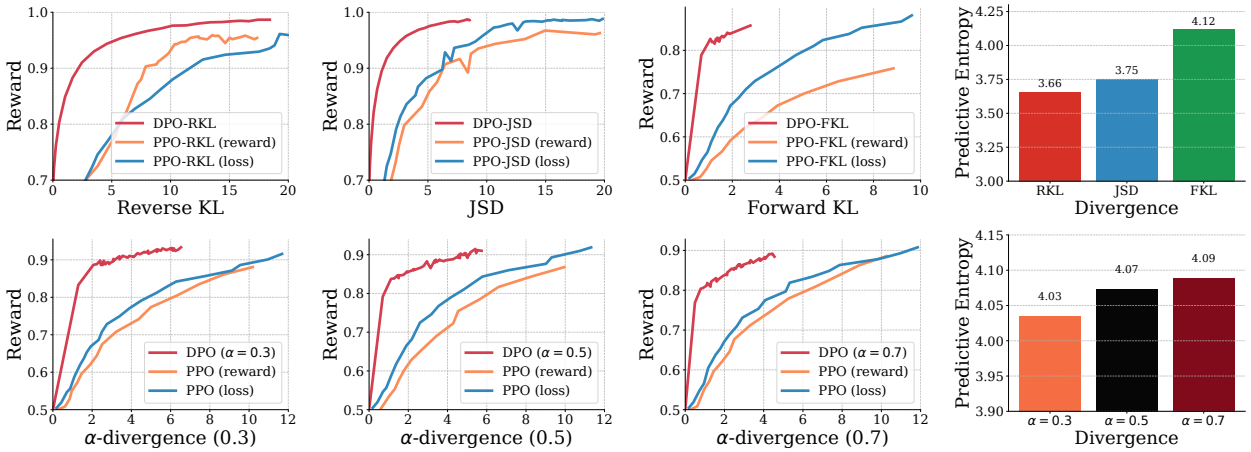


Figure 2.2: Comparisons between DPO with various $f$-divergences and PPO in terms of the frontier of divergence vs reward. To be noted, 'reward' means we add the divergence penalty in the reward, and 'loss' means we add the divergence penalty in the loss.

### *2.6.2 Experiments on IMDB Dataset*

Our initial experiments were performed on the IMDB-sentiment dataset [Maas et al., 2011b] for comparing $f$-DPO against PPO. Following the setup in trlx [CarperAI, 2023], we utilized GPT-2-large [Radford et al., 2019] as our base model for fine-tuning, and the SiEBERT model [Hartmann et al., 2023]—a fine-tuned model of RoBERTa-large [Liu et al., 2019]—was employed for reward computation. For PPO, we explored the divergence coefficient in $\{0.01, 0.03, 0.1, 0.3\}$ for both PPO variants, each using ground-truth rewards. Our PPO implementation is based on the trlx library. Additionally, we adapted the official implementation of DPO with $f$-divergences from Rafailov et al. [2023], setting $\beta$ at 0.1. Please note that PPO utilizes the ground-truth reward during training.

The results are depicted in Figure 2.2, and we also visualize the tradeoff curve in Figure 2.3. Our observations indicate that DPO with $f$-divergences outclasses both PPO implementations in terms of divergence versus reward on the frontier, thus establishing greater divergence-efficiency. Additionally, PPO with a divergence penalty in loss (i.e., PPO (loss)) surpasses PPO with a divergence penalty in reward for both JSD, forward KL and $\alpha$-divergences. This discrepancy arises due to the JSD divergence penalty and forward KL

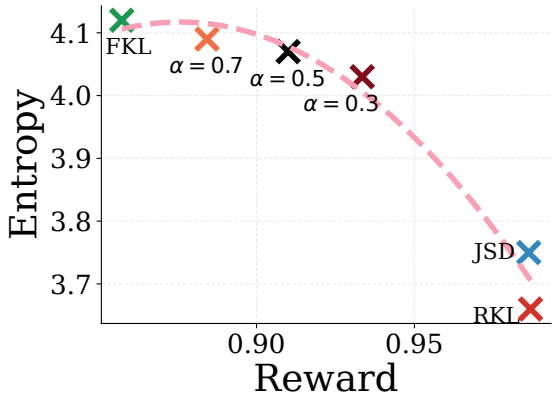


Figure 2.3: The reward and entropy tradeoff of $f$-DPO for different divergences.

penalty fluctuating more significantly than the reverse KL penalty, which consequently introduces instability in learning the value function in PPO. Further experimental results can be found in the Appendix 2.8.1. Finally, we note that reverse KL achieves the lowest predictive entropy due to its mode-seeking property, while Forward KL exhibits the highest predictive entropy. JSD maintains a balance between the two. $\alpha$-divergence interpolates between the JSD and forward KL divergence. This observation aligns with the property of these divergence regularization as well as those in Go et al. [2023], although our framework does not necessitates manual specification of the target distribution or estimation of the normalizing constant under various divergence regularizations.

### *2.6.3 Experiments on Anthropic HH Dataset and MT-bench*

Our next set of experiments was conducted on the Anthropic HH dataset [Bai et al., 2022a]. We adopted the Pythia 2.8B model from Biderman et al. [2023] as our base model. The training configuration follows from Rafailov et al. [2023][1]. The goals of these experiments were to study: 1) how different divergence regularizations impact the trade-off between alignment and diversity in the generated responses, and 2) how $f$-DPO compares to its PPO counterparts. For the first part of the experiments, we utilized automatic metrics for

1. `https://github.com/eric-mitchell/direct-preference-optimization`

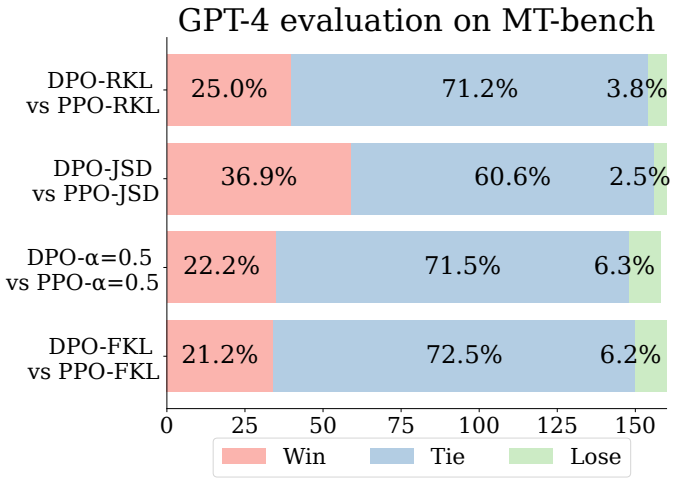


Figure 2.4: MT-Bench comparison between $f$-DPO and PPO under different divergences. The win, tie and lose rates are evaluated based on GPT-4.

evaluation, while for the second part, we relied on the GPT-4 evaluation. For PPO, we adopt the PPO (loss).

The results, in terms of alignment accuracy and diversity, are presented in Table 2.2. To measure diversity, we generated 25 responses using nucleus sampling [Holtzman et al., 2020] with $p = 0.95$ for each prompt in the test set of the Anthropic HH dataset using temperatures of $0.6, 1.0, 1.4$, following Touvron et al. [2023a]. The results for temperatures 0.6 and 1.4 can be found in Appendix 2.8.5. To compute metrics, we employed the predictive entropy, self-bleu [Zhu et al., 2018] and distinct-n [Li et al., 2016]. Consistent with the findings in section 2.6.2, we observed that reverse KL divergence achieves the highest accuracy but the lowest diversity in generation. Adjusting the divergence regularization allows us to trade-off between alignment accuracy and diversity. By tailoring outputs to specific application needs, it offers a customizable balance between accuracy and diversity. This might not only enhance robustness against unfamiliar inputs but also boost user engagement by preventing monotonous interactions.

To compare $f$-DPO with PPO in terms of generation quality, we conducted a pairwise comparison using MT-Bench [Zheng et al., 2023]. MT-Bench is a GPT-4-based evaluation

Table 2.2: Comparison of JSD, RKL, FKL and some $\alpha$-divergences in terms of Alignment and Diversity on Anthropic HH. The ↑ indicates that higher is better, and ↓ means lower is better.

| **Divergences** | **Alignment** | **Diversity** | | | |
|---|---|---|---|---|---|
| | Accuracy (%) ↑ | Entropy ↑ | Self-Bleu ↓ | Distinct-1 ↑ | Distinct-2 ↑ |
| RKL | **67.19** | 12.25 | 0.880 | 0.021 | 0.151 |
| JSD | 66.80 | 12.31 | 0.878 | 0.021 | 0.159 |
| $\alpha = 0.3$ | 59.77 | 12.85 | 0.849 | 0.026 | 0.199 |
| $\alpha = 0.5$ | 61.72 | 12.90 | 0.841 | 0.028 | 0.206 |
| $\alpha = 0.7$ | 57.42 | 12.98 | 0.839 | 0.027 | 0.202 |
| FKL | 54.30 | **13.01** | **0.834** | **0.029** | **0.210** |

benchmark for LLMs that achieves over 80% agreement with human preference judgments on LLM generation quality. MT-Bench includes a series of open-ended questions that assess LLM capabilities in multi-turn conversations and instruction-following, which are often considered important factors for human preference. In accordance with the official MT-Bench implementation [Zheng et al., 2023],[2] we sampled responses with a temperature setting of 0.7 and limited the maximum number of newly generated tokens to 1024. For additional details about MT-bench, we refer readers to the original paper. The GPT-4 evaluation results are provided in Figure 2.4. From these results, it is evident that extending DPO with $f$-divergence achieves performance that is comparable to, and in some cases significantly better than, that of PPO. We've also provided the comparisons between DPO under different $f$-divergence in Section 2.8.6.

### *2.6.4 Experiments on Calibration*

In our last set of experiments, we sought to explore the advantages of divergence efficiency. Our observations revealed that $f$-DPO achieves a smaller divergence than PPO while attaining comparable performance. Previous studies, as referenced by OpenAI [2023a], suggest that RLHF adversely affects the calibration performance of GPT-4. This prompts the question:

2. `https://github.com/lm-sys/FastChat/blob/main/fastchat/llm_judge/gen_model_answer.py`

is there a correlation between the divergence of the base model and that of the finetuned version, and the calibration error? To elucidate this, we presented the following theorem,

**Theorem 2.** *Suppose $\pi_{\boldsymbol{\theta}_1}(\cdot|x)$ and $\pi_{\boldsymbol{\theta}_2}(\cdot|x)$ be two policies. Let $D_f$ be any $f$-divergence such that $f$ is strictly convex.*

$$\mathrm{ECE}(\boldsymbol{\theta}_1) - \mathrm{ECE}(\boldsymbol{\theta}_2) \leq 2\mathbb{E}_X[\psi_f(D_f(\pi_{\boldsymbol{\theta}_1}(\cdot|x), \pi_{\boldsymbol{\theta}_2}(\cdot|x)))]$$

*where $\psi_f$ is a real-valued function such that $\lim_{x\downarrow 0} \psi_f(x) = 0$.*

**Remark 1.** *For JSD, we have* $\mathrm{ECE}(\boldsymbol{\theta}_1) - \mathrm{ECE}(\boldsymbol{\theta}_2) \leq \mathbb{E}_X\left[4\sqrt{2D_{JS}(\pi_{\boldsymbol{\theta}_1}(\cdot|x), \pi_{\boldsymbol{\theta}_2}(\cdot|x))}\right]$.

**Remark 2.** *For KL, we have* $\mathrm{ECE}(\boldsymbol{\theta}_1) - \mathrm{ECE}(\boldsymbol{\theta}_2) \leq \mathbb{E}_X\left[2\sqrt{2D_{KL}(\pi_{\boldsymbol{\theta}_1}(\cdot|x), \pi_{\boldsymbol{\theta}_2}(\cdot|x))}\right]$.

Theorem 2, detailed further in Section 2.8.8, establishes a relationship between Expected Calibration Error (ECE) difference and $f$-divergences. Specifically, the difference in ECE between two models can be bounded by the $f$-divergence. Thus, if the base model exhibits good calibration (i.e., small ECE), a smaller $f$-divergence suggests that the finetuned model is similarly well-calibrated.

To validate the theoretical findings, we conducted experiments to explore the impact of various $f$-divergences on calibration errors. We evaluated calibration error on Anthropic HH dataset [Bai et al., 2022a] with the Pythia 2.8B model from Biderman et al. [2023] as our base model. The model was finetuned using $f$-DPO in the same way as in Section 2.6.3. We treat the task as a binary prediction problem. For predictive probabilities, we use the exponentials of normalized scores where the scores are computed for chosen and rejected strings conditioned on input prompts. The results are shown in Figure 2.5. It is apparent that increased regularization parameters can restrict the extent to which the calibration error can increase. On the other hand, as training progresses, the calibration error increases as well. Similar trends on calibration have been discovered in OpenAI [2023a] as well.

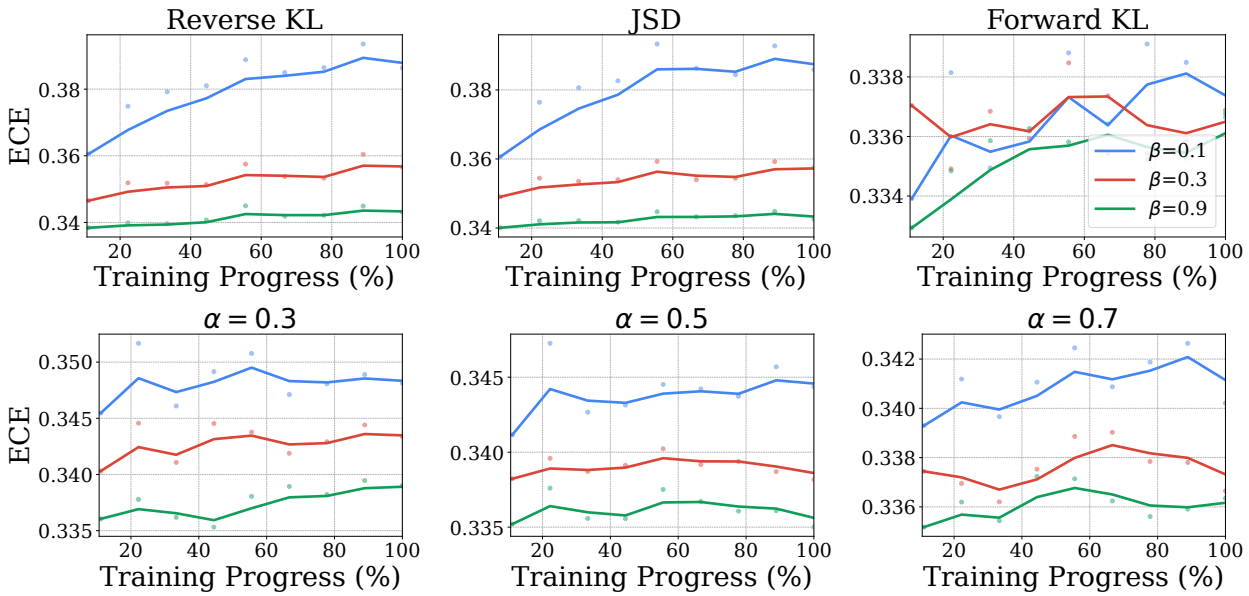


Figure 2.5: Evolution of Expected Calibration Error (ECE) across training steps for three different divergence regularizations: Reverse KL, JSD, Forward KL and $\alpha$-divergence ($\alpha = 0.3, 0.5$ and $0.7$). Each subplot represents the ECE values for varying regularization parameters $\beta = 0.1, 0.3$ and $0.9$ with exponential smoothing.

## 2.7 Conclusion

In this chapter, we introduced a generalized framework for DPO, elegantly incorporating a spectrum of divergence constraints. Our empirical results show that while the reverse KL divergence typically offers superior alignment performance, it compromises generation diversity. By adjusting the divergence regularization, we can achieve a nuanced balance between alignment and generation diversity. Notably, the $f$-DPO framework demonstrates greater divergence efficiency than traditional PPO methods. We further established that the difference in the expected calibration error of two models can be bounded by their divergence, underscoring the advantages of enhanced divergence efficiency. For practitioners, we recommend using JS divergence as the first choice, as it generally generates more diversified responses and is more favored by GPT-4 than reverse KL divergence. Looking forward, we aim to explore the integration of other divergences not well addressed by our present formulation, such as the total variation distance.

## 2.8 Additional Results and Proofs

### *2.8.1 The Cause of Instability in PPO with Divergence Penalty in Rewards*

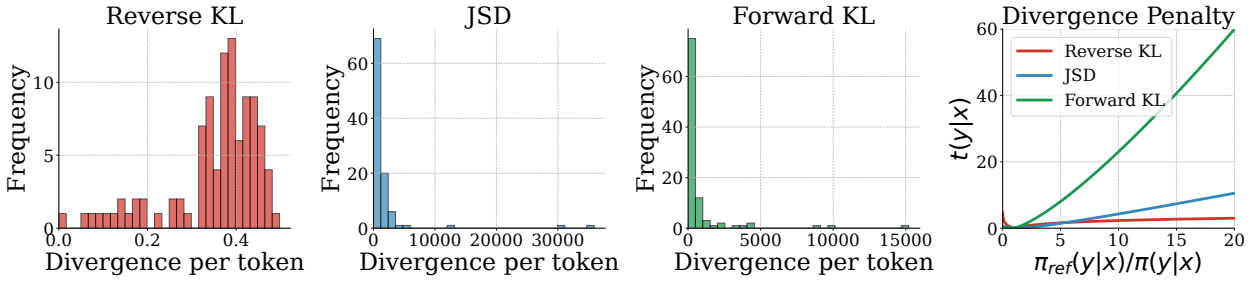


Figure 2.6: Visualization of the divergence penalty for reverse KL, JDS and forward KL.

In RLHF, it is a common practice to impose the KL-divergence as part of the reward function. However, this practice can lead to instability in the optimization process for certain types of divergence measures, such as the forward KL divergence, Jensen-Shannon (JS) divergence, etc. To illustrate this, consider the following reward functions:

$$
\begin{aligned}
r_{\mathrm{RKL}}(y|x) &= r(y|x) + \beta \cdot \log t(y|x), \\
r_{\mathrm{JS}}(y|x) &= r(y|x) - \beta \cdot \left( t(y|x) \log t(y|x) - (t(y|x)+1) \log \left( \frac{t(y|x)+1}{2} \right) \right), \\
r_{\mathrm{FKL}}(y|x) &= r(y|x) - \beta \cdot t(y|x) \log t(y|x),
\end{aligned}
$$

where $t(y|x) = \pi_{\mathrm{ref}}(y|x)/\pi(y|x)$. Figure 2.6 illustrates the curves of the above three divergence penalties as well as the divergence per token computed on the the IMDB dataset during training using PPO with the divergence penalty included in the reward. We oberve that the forward KL divergence penalty will grow much faster than the other two, and the reverse KL divergence penalty grows the slowest. This difference make the reverse KL divergence more numerically stable then the other two and thus makes the learning of the value function much easier.

### 2.8.2 Implementing PPO under $f$-divergence Constraint

To implement PPO under various $f$-divergence, we need to estimate the $f$-divergence using samples. Therefore, an unbiased estimator with low variance is desired. Following Schulman [2020], by denoting $r(x) = p(x)/q(x)$ to be the ratio between two distributions $p$ and $q$, then the following estimator for the $f$-divergence between $p$ and $q$ is unbiased and has low variance,

$$\mathbb{E}_X \left[ f(r(x)) - f'(1)(r(x) - 1) \right].$$

Therefore, we adopt the above estimator for the $f$-divergence when we implement PPO.

### 2.8.3 Additional Discussions on the Choice of Divergences

Different divergences may impose different properties in the optimized solution as we demonstrated through experiments. In pratice, if there is only one model allowed, we would suggest initially considering the Jensen-Shannon (JS) divergence. Our recommendation is based on its ability to yield more diverse responses compared to the commonly used reverse KL divergence. This preference is supported by our findings, where JS divergence can result in a more diversified response as measured by different diversity metrics. Additionally, despite underperforming in reverse KL in reward metrics, it was favored in the GPT-4 evaluation, as illustrated in Figure 2.7 of the appendix. This observation also aligns with recent studies, such as Sun et al. [2023].

To accommodate diverse application needs, another strategy could involve training multiple models, each employing a different divergence, such as JS divergence, $\alpha$ divergence (with $\alpha$ values ranging from 0.1 to 0.9), and forward KL divergence. This approach allows users to experiment and choose the divergence that best aligns with their specific requirements or preferences.

### 2.8.4 Additional Discussions on RKL vs FKL

The distinction between Reverse KL Divergence (RKL) and Forward KL Divergence (FKL) primarily lies in their inherent properties: RKL is mode-seeking, while FKL is mass-covering as we shown in Figure 2.1. This can be discerned from their mathematical formulations. RKL is defined as $E_q[\log(q(x)/p(x))]$, where $q(x)$ is the distribution being optimized, typically representing the language model undergoing fine-tuning. In the RKL scenario, $q(x)$ may assign zero probability to values where $p(x) > 0$. In contrast, FKL is defined as $E_p[\log(p(x)/q(x))]$, which necessitates that $q(x)$ assigns a non-zero probability to all values where $p(x) > 0$. This requirement inherently encourages FKL to promote a distribution $q(x)$ that covers the entire range of $p(x)$, thereby enhancing diversity.

### 2.8.5 Additional Experimental Results

Generation Diversity on Anthropic HH with Different Temperatures

Table 2.3: Comparison of divergences on Anthropic HH with temperature= 0.6.

| **Divergences** | **Self-Bleu** ↓ | **Distinct-1** ↑ | **Distinct-2** ↑ |
|---|---|---|---|
| RKL | 0.8667 | 0.0092 | 0.0615 |
| JSD | 0.8679 | 0.0099 | 0.0662 |
| $\alpha = 0.3$ | 0.8611 | 0.0136 | 0.0899 |
| $\alpha = 0.5$ | 0.8579 | **0.0148** | **0.0950** |
| $\alpha = 0.7$ | 0.8563 | 0.0139 | 0.0905 |
| FKL | **0.8515** | 0.0142 | 0.0926 |

In this section, we report the additional results on evaluating the generation diversity of models trained with different divergences. The temperature were set to be 0.6 and 1.4 following Touvron et al. [2023a]. For each prompt, we sampled 25 responses. The results can be found in Tables 2.3 and 2.4. We found that the pattern is similar to the one observed in the main paper, where we set the temperature to be 1.0.

Table 2.4: Comparison of divergences on Anthropic HH with temperature= 1.4.

| **Divergences** | **Self-Bleu** ↓ | **Distinct-1** ↑ | **Distinct-2** ↑ |
|---|---|---|---|
| RKL | 0.7975 | 0.0973 | 0.6574 |
| JSD | 0.7995 | 0.1025 | 0.6439 |
| $\alpha = 0.3$ | 0.7759 | 0.1107 | 0.6952 |
| $\alpha = 0.5$ | 0.7603 | 0.1101 | 0.6692 |
| $\alpha = 0.7$ | **0.7537** | 0.1151 | 0.6659 |
| FKL | 0.7566 | **0.1233** | **0.7082** |

### *2.8.6 GPT-4 Evaluations on MT-Bench for DPO with Different Divergences*

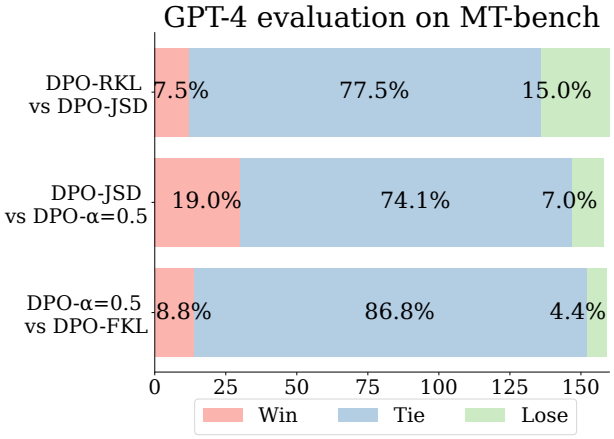


Figure 2.7: Comparing DPO with different divergence regularizations using GPT-4 on MT-Bench.

We further provide a comparison between DPO using different divergences, with GPT-4 serving as the referee on MT-bench. We observe that while DPO with reverse KL outperforms DPO with JSD in terms of alignment accuracy on Anthropic HH, it underperforms when evaluated by GPT-4. Additionally, DPO with JSD performs better than DPO with $\alpha$-divergence. Lastly, DPO with $\alpha$-divergence performs slightly better than DPO with forward KL divergence. These results largely align with our expectations.

### *2.8.7 Proofs*

**Theorem 1.** *If* $\pi_{\text{ref}}(y|x) > 0$ *for any valid* $x$ *and* $f'$ *is invertible with* $0 \notin dom(f')$*, the*

*reward class that is consistent with the Bradley-Terry model can be reparameterized using the policy model $\pi(y|x)$ and a reference model $\pi_{\text{ref}}(y|x)$ as*

$$r(y|x) = \beta f'\left(\frac{\pi^\star(y|x)}{\pi_{\text{ref}}(y|x)}\right) + \text{const.} \tag{2.1}$$

*Proof.* The Karush-Kuhn-Tucker (KKT) conditions for the given optimization problem can be stated as follows:

1. **Stationarity Condition**:

   This condition requires that the gradient of the Lagrangian with respect to the primal variables be zero:

   $$\nabla_{\pi(y|x)}\mathcal{L}(\pi, \lambda, \alpha) = 0, \ \forall y.$$

   By setting the derivative of the Lagrangian with respect to $\pi(y|x)$ to zero, we obtain:

   $$r(y|x) - \beta\frac{\partial}{\partial\pi(y|x)}\mathbb{E}_{\pi_{\text{ref}}}\left[f\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right] - \lambda + \alpha(y) = 0, \ \forall y.$$

2. **Primal Feasibility**: This condition requires that the solution satisfy the original constraints of the problem:

   $$\sum_y \pi(y|x) = 1, \quad \text{and} \quad \pi(y|x) \geq 0 \ \forall y.$$

3. **Dual Feasibility**: This condition requires that the Lagrange multipliers corresponding to inequality constraints are non-negative:

   $$\alpha(y) \geq 0, \ \forall y.$$

4. **Complementary Slackness**: This condition requires that for each inequality con-

straint, either the constraint is satisfied with equality, or the corresponding Lagrange multiplier is zero:

$$\alpha(y)\pi(y|x) = 0, \ \forall y.$$

In this context, $\pi(y|x)$ is the primal variable we're optimizing, $\lambda$ is the Lagrange multiplier for the equality constraint, and $\alpha(y)$ are the Lagrange multipliers for the inequality constraints.

To derive the final solution to the given problem, we first use the stationarity condition to obtain an equation that relates $\pi(y|x)$, $r(y|x)$, $f$, $\lambda$, and $\alpha(y)$. This will involve taking the derivative of $f\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right)$ with respect to $\pi(y|x)$, which will depend on the specific form of the function $f$.

We denote the derivative of $f$ with respect to its argument as $f'$, and that the derivative of $r(y|x)$ with respect to $\pi(y|x)$ is zero (since $r(y|x)$ does not explicitly depend on $\pi(y|x)$), we get:

$$r(y|x) - \beta f'\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right) - \lambda + \alpha(y) = 0.$$

We can solve this for $\pi(y|x)$, assuming that the inverse of $f'$ exists:

$$\pi(y|x) = \pi_{\text{ref}}(y|x) f'^{-1}\left(\frac{r(y|x) - \lambda + \alpha(y)}{\beta}\right).$$

However, we have to note that this is under the assumptions that $f'$ is invertible, and that the inverse maps the argument into a domain where the original function $f$ is defined and differentiable.

Next, we can use the primal feasibility condition to get an equation that will help determine the values of $\lambda$ and $\alpha(y)$. Substituting the expression for $\pi(y|x)$ into the constraint

$\sum_y \pi(y|x) = 1$, we obtain an equation that can be solved for $\lambda$:

$$\sum_y \pi_{\text{ref}}(y|x) f'^{-1}\left(\frac{r(y|x) - \lambda + \alpha(y)}{\beta}\right) = 1.$$

This equation is likely to be nonlinear and might need numerical methods to solve for $\lambda$ and $\alpha(y)$. Also, we need to ensure $\pi(y|x) \geq 0$ for all $y$, which could place further restrictions on the possible values of $\lambda$ and $\alpha(y)$.

Finally, the complementary slackness condition $\alpha(y)\pi(y|x) = 0$ for all $y$ will eliminate some solutions, because for each $y$, either $\alpha(y) = 0$ or $\pi(y|x) = 0$ must hold.

Therefore, we can write out the reward function as a function of the policy, i.e.,

$$r(y|x) = \beta f'\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right) + \lambda - \alpha(y).$$

The complementary slackness requires that

$$\pi(y|x)\alpha(y) = 0 \ \ \forall y.$$

Hence, for those function $f$ that $0 \notin \text{dom}(f')$ with the assumption that $\pi_{\text{ref}}(y|x) > 0$ almost surely, we must have $\alpha(y) = 0 \ \forall y$. In particular, the reverse KL, forward KL and JS divergences are among this category, See Table 2.1. Thus, we can simplify the reward function as,

$$r(y|x) = \beta f'\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right) + \lambda,$$

where $\lambda$ depends only on $x$ and thus can be treated as an additive constant, which will be cancelled out in the BT model.

□

### *2.8.8 On Calibration and $f$-divergence*

To measure calibration, we adopt the definition of Expected Calibration Error (ECE) [Guo et al., 2017]. Specifically, for any policy $\pi_{\boldsymbol{\theta}}(\cdot|x)$, define $\hat{P}_{\pi_{\boldsymbol{\theta}}}(x)$ as

$$\hat{P}_{\pi_{\boldsymbol{\theta}}}(x) = \pi_{\boldsymbol{\theta}}(\hat{y}|x)$$

where $\hat{y}$ is the predicted label of $x$. Given a policy $\pi_{\boldsymbol{\theta}}$, $\hat{y}$ is sampled with probability $\pi_{\boldsymbol{\theta}}(\hat{y}|x)$.

Then ECE can be defined as follows:

$$\mathrm{ECE}(\boldsymbol{\theta}) = \mathbb{E}_{\hat{P}_{\pi_{\boldsymbol{\theta}}}}[|\mathbb{P}(\hat{Y} = Y|\hat{P}_{\pi_{\boldsymbol{\theta}}} = p) - p|]$$

**Remark 3.** *Here, we consider our policy to be stochastic. This is different from the definition in [Guo et al., 2017] where $\hat{y}$ is chosen to be the one with the highest probabilities.*

The following theorem bound the differences of ECE by $f$ divergences of policies.

**Theorem 2.** *Suppose $\pi_{\boldsymbol{\theta}_1}(\cdot|x)$ and $\pi_{\boldsymbol{\theta}_2}(\cdot|x)$ be two policies. Let $D_f$ be any $f$-divergence such that $f$ is strictly convex.*

$$\mathrm{ECE}(\boldsymbol{\theta}_1) - \mathrm{ECE}(\boldsymbol{\theta}_2) \leq 2\mathbb{E}_X[\psi_f(D_f(\pi_{\boldsymbol{\theta}_1}(\cdot|x), \pi_{\boldsymbol{\theta}_2}(\cdot|x)))]$$

*where $\psi_f$ is a real-valued function such that $\lim_{x\downarrow 0} \psi_f(x) = 0$.*

*Proof.* By the tower rule, we know that

$$\begin{aligned}
\mathrm{ECE}(\boldsymbol{\theta}) &= \mathbb{E}_X\left[\mathbb{E}_{\hat{P}_{\pi_{\boldsymbol{\theta}}}|X}\left[\left|\mathbb{P}(\hat{Y}=Y|\hat{P}_{\pi_{\boldsymbol{\theta}}}=p, X=x)-p\right|\right]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_{\hat{P}_{\pi_{\boldsymbol{\theta}}}(X)}\left[\left|\mathbb{P}(\hat{Y}=Y|\hat{P}_{\pi_{\boldsymbol{\theta}}}=p, X=x)-p\right|\right]\right] \\
&= \mathbb{E}_X\left[\sum_{\hat{y}}\pi_{\boldsymbol{\theta}}(\hat{y}|x)\left|\mathbb{P}\left(Y=\hat{y}\middle|\hat{P}_{\pi_{\boldsymbol{\theta}}}=\pi_{\boldsymbol{\theta}}(\hat{y}|x), X=x\right)-\pi_{\boldsymbol{\theta}}(\hat{y}|x)\right|\right] \\
&= \mathbb{E}_X\left[\sum_{\hat{y}}\pi_{\boldsymbol{\theta}}(\hat{y}|x)\left|\mathbb{P}(Y=\hat{y}|X=x)-\pi_{\boldsymbol{\theta}}(\hat{y}|x)\right|\right]
\end{aligned}$$

Let $\pi(y|x)$ be the conditional distribution of ground truth. Then,

$$\mathrm{ECE}(\boldsymbol{\theta}) = \mathbb{E}_X\left[\left\langle|\pi(\cdot|x)-\pi_{\boldsymbol{\theta}}(\cdot|x)|, \pi_{\boldsymbol{\theta}}(\cdot|x)\right\rangle\right]$$

Here, both $|\pi(\cdot|x)-\pi_{\boldsymbol{\theta}}(\cdot|x)|$ and $\pi_{\boldsymbol{\theta}}(\cdot|x)$ are vectors where each entry of $|\pi(\cdot|x)-\pi_{\boldsymbol{\theta}}(\cdot|x)|$ is the absolute value of the corresponding entry of $\pi(\cdot|x)-\pi_{\boldsymbol{\theta}}(\cdot|x)$. Now, let's compare the calibration errors of two models $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,

$$\begin{aligned}
&\mathrm{ECE}(\boldsymbol{\theta}_1) - \mathrm{ECE}(\boldsymbol{\theta}_2) \\
&\leq \mathbb{E}_X\left[\left\langle|\pi(\cdot|x)-\pi_{\boldsymbol{\theta}_1}(\cdot|x)|, \pi_{\boldsymbol{\theta}_1}(\cdot|x)\right\rangle - \left\langle|\pi(\cdot|x)-\pi_{\boldsymbol{\theta}_2}(\cdot|x)|, \pi_{\boldsymbol{\theta}_2}(\cdot|x)\right\rangle\right] \\
&\leq \mathbb{E}_X\left[\left\langle\frac{\pi_{\boldsymbol{\theta}_1}(\cdot|x)+\pi_{\boldsymbol{\theta}_2}(\cdot|x)+|\pi_{\boldsymbol{\theta}}(\cdot|x)-\pi_{\boldsymbol{\theta}_1}(\cdot|x)|+|\pi_{\boldsymbol{\theta}}(\cdot|x)-\pi_{\boldsymbol{\theta}_2}(\cdot|x)|}{2}, |\pi_{\boldsymbol{\theta}_1}(\cdot|x)-\pi_{\boldsymbol{\theta}_2}(\cdot|x)|\right\rangle\right]
\end{aligned}$$

By Holder's inequality, we have that

$$\begin{aligned}
&\mathrm{ECE}(\boldsymbol{\theta}_1) - \mathrm{ECE}(\boldsymbol{\theta}_2) \\
&\leq \mathbb{E}_X\left[\|\pi_{\boldsymbol{\theta}_1}(\cdot|x)-\pi_{\boldsymbol{\theta}_2}(\cdot|x)\|_1 \cdot \max\frac{\pi_{\boldsymbol{\theta}_1}(\cdot|x)+\pi_{\boldsymbol{\theta}_2}(\cdot|x)+|\pi_{\boldsymbol{\theta}}(\cdot|x)-\pi_{\boldsymbol{\theta}_1}(\cdot|x)|+|\pi_{\boldsymbol{\theta}}(\cdot|x)-\pi_{\boldsymbol{\theta}_2}(\cdot|x)|}{2}\right]
\end{aligned}$$

Let

$$m(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, x) = \max \frac{\pi_{\boldsymbol{\theta}_1}(\cdot|x) + \pi_{\boldsymbol{\theta}_2}(\cdot|x) + |\pi_{\boldsymbol{\theta}}(\cdot|x) - \pi_{\boldsymbol{\theta}_1}(\cdot|x)| + |\pi_{\boldsymbol{\theta}}(\cdot|x) - \pi_{\boldsymbol{\theta}_2}(\cdot|x)|}{2}$$

then

$$\begin{aligned}
\mathrm{ECE}(\boldsymbol{\theta}_1) - \mathrm{ECE}(\boldsymbol{\theta}_2) &\leq \mathbb{E}_X[\|\pi_{\boldsymbol{\theta}_1}(\cdot|x) - \pi_{\boldsymbol{\theta}_2}(\cdot|x)\|_1 m(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, x)] \\
&= \mathbb{E}_X[2\delta(\pi_{\boldsymbol{\theta}_1}(\cdot|x), \pi_{\boldsymbol{\theta}_2}(\cdot|x)) m(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, x)] \\
&\leq \mathbb{E}_X\left[4\delta(\pi_{\boldsymbol{\theta}_1}(\cdot|x), \pi_{\boldsymbol{\theta}_2}(\cdot|x))\right] \\
&\leq \mathbb{E}_X\left[2\psi_f(D_f(\pi_{\boldsymbol{\theta}_1}(\cdot|x), \pi_{\boldsymbol{\theta}_2}(\cdot|x)))\right]
\end{aligned}$$

where $\delta$ is the total variation distance, $D_f$ is any $f$ divergence such that $f$ is strictly convex and $\psi_f$ is a real-valued function such that $\lim_{x \downarrow 0} \psi_f(x) = 0$. See [Sason and Verdú, 2016] for more details on the last inequality. □

# CHAPTER 3

# PREFERENCE OPTIMIZATION WITH MULTI-SAMPLE COMPARISONS

## 3.1 Introduction

The $f$-dpo introduced in the prior chapter addresses the issue of entropy collapse. However, controlling generative distribution via entropy is too crude and is not controllable enough. Recent advancements in generative models, particularly large language models (LLMs) and diffusion models, have been driven by extensive pretraining on large datasets followed by post-training. However, current post-training methods such as reinforcement learning from human feedback (RLHF) and direct alignment from preference methods (DAP) primarily utilize single-sample comparisons. These approaches often fail to capture critical characteristics such as generative diversity and bias, which are more accurately assessed through multiple samples. To address these limitations, we introduce a novel approach that extends post-training to include multi-sample comparisons. To achieve this, we propose Multi-sample Direct Preference Optimization (mDPO) and Multi-sample Identity Preference Optimization (mIPO), which enables more precise control in the final generative distribution. These methods improve traditional DAP methods by focusing on group-wise characteristics. Empirically, we demonstrate that multi-sample comparison is more effective in optimizing collective characteristics (e.g., diversity and bias) for generative models than single-sample comparison. Additionally, our findings suggest that multi-sample comparisons provide a more robust optimization framework, particularly for dataset with label noise.

## 3.2 Controllable Finetuning of Generative Models

Generative models, particularly large language models (LLMs) [Achiam et al., 2023, Meta AI, 2024, Bai et al., 2023, Bi et al., 2024, Gemini et al., 2023, Anthropic, 2024] and diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020, Rombach et al., 2022, Podell et al., 2023], hold the tremendous promise to transform numerous industries by automating complex tasks, enhancing creativity, and personalizing user experiences at an unprecedented scale [Eloundou et al., 2023]. These models achieve their capabilities through extensive pretraining on large-scale datasets, followed by post-training to unlock the capabilities (i.e., the superficial alignment hypothesis) [Wei et al., 2022, Tay et al., 2022, Zhou et al., 2024a]. For post-training stage, reinforcement learning from human feedbacks (RLHF) and direct alignment from preference methods (DAP) have been crucial for the success of both LLMs [Ouyang et al., 2022, Rafailov et al., 2024, Azar et al., 2024] and diffusion models [Black et al., 2023, Wallace et al., 2023, Yang et al., 2024a].

Despite these advancements, current post-training approaches predominantly focus on single-sample comparisons, which fail to capture characteristics better assessed through distributions of samples, such as creativity and bias. Evaluating a model's creativity/consistency or detecting biases requires analyzing the variability and diversity across multiple outputs, not just individual ones. For example, while LLMs are proficient in crafting narratives, they often show limitations in generating a diverse representation of genres [Patel et al., 2024, Wang et al., 2024d]. Additionally, these models tend to have lower entropy in their predictive distributions after post-training, leading to limited generative diversity [Mohammadi, 2024, Wang et al., 2023a, Wiher et al., 2022, Khalifa et al., 2020].

Consider the task of generating a random integer between 0 and 10. Ideally, the model should not prefer any particular number. However, Zhang et al. [2024b] show that existing models often bias towards certain numbers over the others. Similar issues are observed in diffusion models, which may display biases in the generated outputs based on factors like

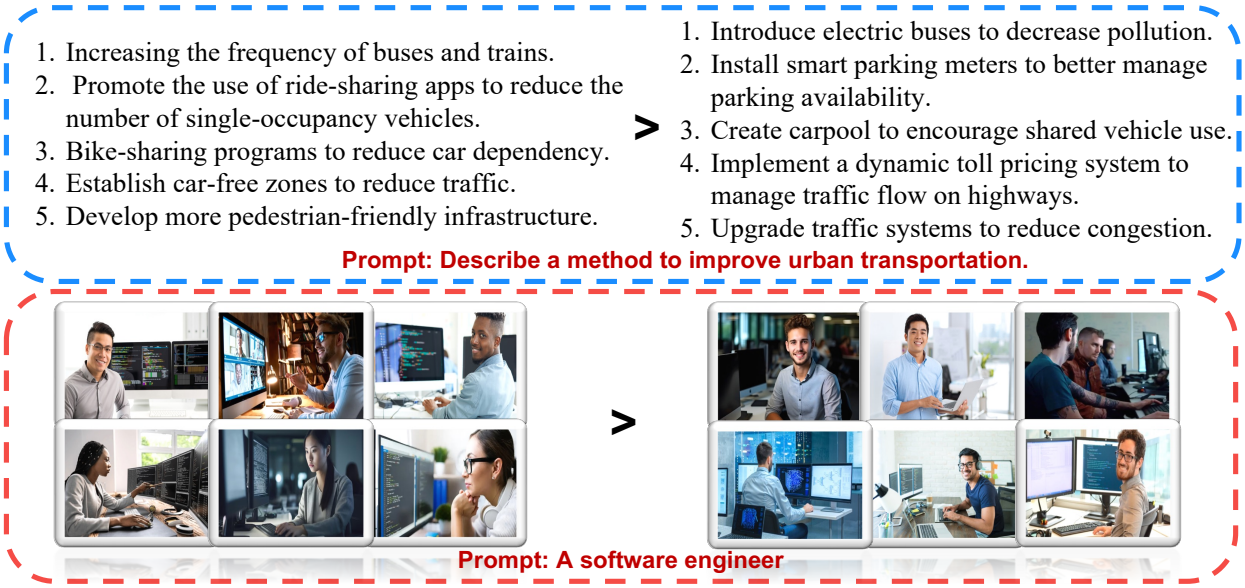


Figure 3.1: **Top**: Diversity of responses from two groups for improving urban transportation. The left group provides a broader range of approaches, including public transit and infrastructure improvements. The right group focuses more narrowly on specific technological and management solutions. **Bottom**: Bias in images from two groups. The left group displays a more balanced representation of race and gender, while the right group predominantly features males due to stereotypes.

gender or race [Luccioni et al., 2023, Chen et al., 2024b]. Inconsistencies in generation is also a crucial issue that needs to be addressed to make models more reliable [Liu et al., 2023, Bubeck et al., 2023]. The aforementioned failures cannot be captured by a single sample; instead, they are distributional issues (see figure 3.1 for an illustration). To address the limitations, we extend the post-training paradigm to multi-sample comparisons. This approach involves evaluating the model's performance over distributions of samples rather than individual samples. By curating groups of responses and assessing their collective characteristics, we can better align the model's outputs with desired distributional properties.

In this chapter, we introduce Multi-sample Direct Preference Optimization (mDPO) and Multi-sample Identity Preference Optimization (mIPO), which are extensions of the prior

DAP methods DPO [Rafailov et al., 2024] and IPO [Azar et al., 2024][1]. Unlike their predecessors, which rely on single-sample comparisons, mDPO and mIPO utilize multi-sample comparisons to better capture group-wise or distributional characteristics. Our experiments involve fine-tuning language models to generate random numbers in a calibrated manner and enhancing the diversity of genres in creative story writing. Additionally, for diffusion models, we demonstrate a significant reduction in gender and race biases compared to the traditional single-sample comparison approach. Furthermore, we present evidence that multi-sample comparison offers a more robust optimization method in the presence of label noise in preference data, such as synthetic data that lacks human labeling but possesses knowledge of the overall quality between two models.

## 3.3 Related Works

**Reinforcement Learning from Human Feedback (RLHF)** has been pivotal in advancing the capabilities of large language models [Ouyang et al., 2022, Bai et al., 2022a] and, more broadly, in agent alignment [Christiano et al., 2017, Leike et al., 2018]. RLHF was initially introduced by Christiano et al. [2017] to tackle classic reinforcement learning tasks such as those found in MuJoCo. The work by Ouyang et al. [2022] marked the first significant application of RLHF in aligning language models to follow human instructions, thereby establishing a foundation for subsequent models like ChatGPT [Achiam et al., 2023]. To address the complexities and instabilities associated with RL methods, Direct Alignment from Preference (DAP) methods [Rafailov et al., 2024, Azar et al., 2024, Dong et al., 2023, Yuan et al., 2023, Zhao et al., 2023b] have been proposed. Notably, Rafailov et al. [2024] introduced Direct Preference Optimization (DPO), which reframes the RL problem into a supervised learning problem by deriving an analytical relationship between policy and reward. To counteract potential overoptimization issues in DPO, Azar et al. [2024] proposed Identity

1. Our framework can also be extended to other preference optimization algorithms.

Preference Optimization (IPO), which employs an $\ell_2$ loss to regress the margin towards a predefined threshold. In the context of diffusion models, Black et al. [2023] introduced Denoising Diffusion Policy Optimization (DDPO) to optimize diffusion models using specific reward functions. Wallace et al. [2023] expanded the application of DPO from language domains to image domains. However, these methods predominantly focus on single-sample comparisons and may not adequately capture the collective characteristics of output distributions.

**Distributional difference** [Zhong et al., 2022] describes the difference between two or more distributions, which is a generalization of single-sample comparison. While many characteristics can be judged from a single sample, some properties can only be deduced from multiple samples, such as diversity and bias [Santurkar et al., 2023, Chen et al., 2024b, Zhang et al., 2024c, Zhou et al., 2024b, Mohammadi, 2024, Wang et al., 2023a, Go et al., 2023]. To measure the distributional difference, Zhong et al. [2022] proposed a hypothesis-verifier framework to summarize the difference between two corpora. Dunlap et al. [2024] further extended this framework to the image domain to capture the difference between two sets of images. More recently, Zhong et al. [2023] considered the problem of distributional difference in a goal-driven setting. Melnyk et al. [2024] addressed the problem of distributional alignment for language models using optimal transport, aiming to optimize the chosen responses across all prompts jointly. In contrast, our work addresses an orthogonal problem by focusing on optimizing the distributional preference under the same prompt.

## 3.4 Preliminaries

**Supervised Finetuning.** After the language model has been pretrained on extensive datasets, the next step is typically supervised finetuning (SFT). This process aims to refine the model's predictions, preparing it for subsequent alignment stages. During SFT, the pretrained language model is finetuned using a supervised dataset $\mathcal{D} = \{(x, y)_i\}_{i=1}^N$, where

$x$ represents the input and $y$ denotes the target. The objective of SFT is to minimize the negative log-likelihood,

$$\mathcal{L}_{\text{SFT}}(\pi_{\boldsymbol{\theta}}, \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[ -\log \pi_{\boldsymbol{\theta}}(y|x) \right].$$

**Direct Alignment from Preference.** For single-sample comparison setting, the dataset consists of $\{(x, y_w, y_l)_i\}_{i=1}^N$, where $x$ denotes the prompt or the input to the model, $y_w$ is the preferred response over $y_l$. Under the Bradley-Terry model, the likelihood of $y_w$ is preferred to $y_l$ is computed by

$$p(y_w \succcurlyeq y_l|x) = \sigma(r(y_w, x) - r(y_l, x)) = \frac{\exp(r(y_w, x))}{\exp(r(y_w, x)) + \exp(r(y_l, x))},$$

where $r(y, x)$ is the scalar reward of answering $y$ when given $x$. Direct preference optimization (DPO) [Rafailov et al., 2024] establishes an analytical form between the reward function and the policy or the language model, i.e., the language model represents an implicit reward. Specifically, under DPO, the implicit reward can be computed by

$$r_{\boldsymbol{\theta}}(y, x) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\text{ref}}(y|x)} + \text{const}(x).$$

Using the implicit reward, DPO optimizes the policy $\pi_{\boldsymbol{\theta}}$ by minimizing the negative log-likelihood on the offline preference dataset,

$$\mathcal{L}_{\text{DPO}}(\pi_{\boldsymbol{\theta}}, \mathcal{D}) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[ -\log \sigma \left( r_{\boldsymbol{\theta}}(y_w, x) - r_{\boldsymbol{\theta}}(y_l, x) \right) \right].$$

$$\mathcal{L}_{\text{DPO}}(\pi_{\boldsymbol{\theta}}, \mathcal{D}) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[ -\log \sigma \left( \beta \log \frac{\pi_{\boldsymbol{\theta}}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\boldsymbol{\theta}}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right].$$

Although the Bradley-Terry model is based on the assumption of Gumbel noise in the reward function, a choice that is well-suited for discrete random variables, it has certain limitations,

such as overfitting, as noted by Azar et al. [2024]. Specifically, Identity Preference Optimization (IPO) [Azar et al., 2024] addresses these limitations by bypassing the Bradley-Terry model for preference modeling. IPO aims to mitigate the issue of "overfitting" in preference datasets by replacing the sigmoid function with the squared distance,

$$\mathcal{L}_{\text{IPO}}(\pi_{\boldsymbol{\theta}}, \mathcal{D}) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \frac{\pi_{\boldsymbol{\theta}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)} - \log \frac{\pi_{\boldsymbol{\theta}}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \frac{\tau^{-1}}{2}\right]^2,$$

which can be interpreted as regressing the difference between the log-likelihood ratio to $\tau^{-1}/2$.

## 3.5 Method

The singleton preference (i.e., the marginal distribution) may not generalize well to the distributional preference (i.e., the joint distribution) [Melnyk et al., 2024]. For most of the cases, when given a prompt $x$, there is indeed a preference between two responses $y_1$ and $y_2$. However, there are also cases where the preference cannot be assigned purely based on two singletons. Consider the cases where we prompt the diffusion models to generate an image of "a software engineer". For such cases, there is no preference between "a female software engineer" and "a male software engineer". Similarly, if a language model is prompted to generate an integer randomly and uniformly from $\{0, 1, ..., 9\}$, there is also no preference between generating any two specific numbers, e.g., 3 or 5. Nonetheless, preferences can exist at a **multi-sample or joint distribution level**.

### *3.5.1 Multi-sample Preference Optimization*

Thus, we generalize the preference model from singleton comparison to multi-sample comparison, where the preference is assigned at the multi-sample level instead of on a singleton-level. Let $\mathcal{G}_w$ and $\mathcal{G}_l$ denote the multiple samples sampled from the model's conditional distribu-

tion[2] when given the prompt $x$, then the likelihood of $\mathcal{G}_w$ is preferred over $\mathcal{G}_l$ is defined as

$$p(\mathcal{G}_w \succcurlyeq \mathcal{G}_l|x) = \Phi(r(\mathcal{G}_w, x) - r(\mathcal{G}_l, x)),$$

where the function $\Phi$ follows the definition in the IPO paper [Azar et al., 2024], which can be e.g., the sigmoid function (recovers the Bradley-Terry model). Given the reward function $r(\cdot, \cdot)$, the goal of RLHF is to maximize the reward under reverse KL constraints. This can be captured by the following optimization objective,

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{G}\sim\pi_{\boldsymbol{\theta}}|x, x\sim p_x} \left[r(\mathcal{G}, x) - \beta \cdot \mathrm{KL}(\pi_{\boldsymbol{\theta}}(\cdot|x)||\pi_{\mathrm{ref}}(\cdot|x))\right].$$

Generalizing the DPO's derivation, we obtain the following equation for the reward function,

$$r(\mathcal{G}, x) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathcal{G}|x)}{\pi_{\mathrm{ref}}(\mathcal{G}|x)} + \mathrm{const}(x).$$

where $\mathrm{const}(x)$ is some constant that only depends $x$, and $\beta$ is the coefficient of the KL regularization. For a given dataset, $\{(\mathcal{G}_w, \mathcal{G}_l, x)_i\}_{i=1}^N$, where $\mathcal{G} = \{y_j\}_{j=1}^k$ with $y_j \overset{\text{i.i.d}}{\sim} \pi(\cdot|x)$, we can expand the likelihood by $\pi_\theta(\mathcal{G}|x) = \prod_{y\in\mathcal{G}} \pi_{\boldsymbol{\theta}}(y|x)$. To avoid the effect of the size of $\mathcal{G}$, we consider using the geometric mean instead, $\pi_\theta(\mathcal{G}|x) = (\prod_{y\in\mathcal{G}} \pi_{\boldsymbol{\theta}}(y|x))^{1/|\mathcal{G}|}$. Thus the objective of multi-sample DPO becomes (where $\Phi(z) = \sigma(z)$),

$$\mathcal{L}_{\mathrm{mDPO}} = \mathbb{E}_{(x,\mathcal{G}_w,\mathcal{G}_l)\sim\mathcal{D}} \left[-\log\sigma\left(\beta\mathbb{E}_{y_w\sim\mathcal{G}_w}\left[\log\frac{\pi_{\boldsymbol{\theta}}(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)}\right] - \beta\mathbb{E}_{y_l\sim\mathcal{G}_l}\left[\log\frac{\pi_{\boldsymbol{\theta}}(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right]\right)\right].$$

Similarly, for the IPO variant (where $\Phi(z) = (z - \tau^{-1}/2)^2$), we have

$$\mathcal{L}_{\mathrm{mIPO}} = \mathbb{E}_{(x,\mathcal{G}_w,\mathcal{G}_l)\sim\mathcal{D}} \left[\left(\mathbb{E}_{y_w\sim\mathcal{G}_w}\left[\log\frac{\pi_{\boldsymbol{\theta}}(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)}\right] - \mathbb{E}_{y_l\sim\mathcal{G}_l}\left[\log\frac{\pi_{\boldsymbol{\theta}}(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right] - \frac{\tau^{-1}}{2}\right)^2\right].$$

2. The definition of a group $\mathcal{G}$ can be generalized to distributions. Thus, $\mathcal{G}$ is equivalent to the predictive distribution of the language model.

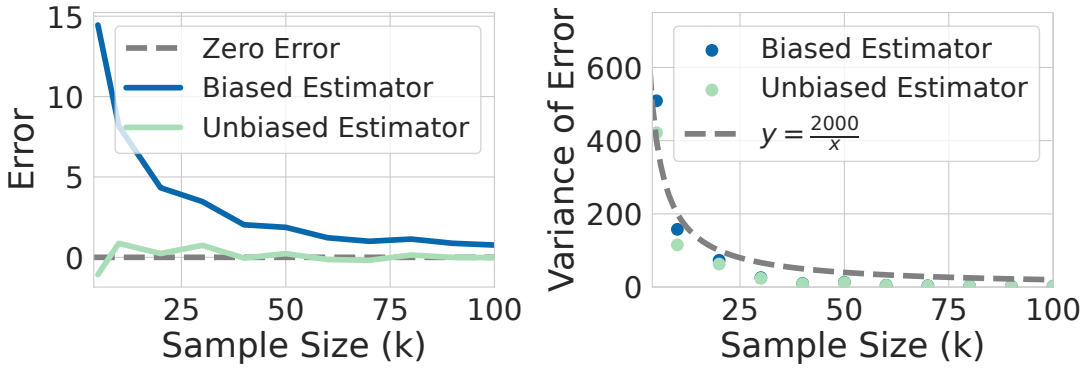


Figure 3.2: Biased estimator vs. Unbiased estimator.

The main difference from their original objectives is that the implicit reward for a single sample is replaced with the averaged implicit reward for a group of samples. We use the notation $y \sim \mathcal{G}$ to account for the general case where $\mathcal{G}$ represents a distribution rather than merely a finite collection of samples.

### *3.5.2 Stochastic Estimator for Efficient Optimization*

Our focus is on finetuning large generative models, and thus a scalable estimator for the gradient / objective is necessary to make the algorithm practically useful. Deriving a mini-batch (potentially unbiased and low-variance) estimator for mDPO is challenging due to the non-linearity of the sigmoid function. However, for mIPO, this task is more tractable by expanding the square function. This is equivalent to computing the unbiased estimator for the following objective, $\ell = \left(\mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{x\sim q}[f(x)] - c\right)^2$. The unbiased estimator for mIPO is stated in the following result.

**Proposition 1.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function, and let $p$ and $q$ be probability distributions on $\mathcal{X}$. Define $\ell = (\mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{x\sim q}[f(x)] - c)^2$, where $c$ is a constant. Let*

*$x_1^p, \ldots, x_n^p$ be i.i.d. samples from $p$, and $x_1^q, \ldots, x_m^q$ be i.i.d. samples from $q$. Then,*

$$\hat{\ell} = \left( \frac{1}{n} \sum_{i=1}^{n} f(x_i^p) - \frac{1}{m} \sum_{j=1}^{m} f(x_j^q) - c \right)^2 - \left( \frac{\hat{\sigma}_p^2}{n} + \frac{\hat{\sigma}_q^2}{m} \right)$$

*is an unbiased estimator of $\ell$, where $\hat{\sigma}_p^2$ and $\hat{\sigma}_q^2$ are the sample variances of $f(x)$ under $p$ and $q$.*

Therefore, to optimize the mIPO objective, we can sample mini-batches similar to IPO. Each mini-batch can consist of multiple responses from both $\mathcal{G}_w$ and $\mathcal{G}_l$. While increasing the number of samples from each group reduces the variance of the estimated gradient, it also raises computational and memory costs. In summary, the mIPO objective is thus

$$\mathbb{E}_{(x,\{y_{w,i}\}_{i=1}^k,\{y_{l,i}\}_{i=1}^k)\sim\mathcal{D}} \left[ \left( \frac{1}{k} \sum_{i=1}^{k} \log \frac{\pi_{\boldsymbol{\theta}}(y_{w,i}|x)}{\pi_{\text{ref}}(y_{w,i}|x)} - \frac{1}{k} \sum_{j=1}^{k} \log \frac{\pi_{\boldsymbol{\theta}}(y_{l,j}|x)}{\pi_{\text{ref}}(y_{l,j}|x)} - \frac{\tau^{-1}}{2} \right)^2 - \frac{\hat{\sigma}_w^2}{k} - \frac{\hat{\sigma}_l^2}{k} \right].$$

For mDPO, we consider the following objective for multi-sample comparison, which may exhibit low variance (for larger values of $k$) but is biased,

$$\mathbb{E}_{(x,\{y_{w,i}\}_{i=1}^k,\{y_{l,i}\}_{i=1}^k)\sim\mathcal{D}} \left[ -\log \sigma \left( \frac{\beta}{k} \sum_{i=1}^{k} \log \frac{\pi_{\boldsymbol{\theta}}(y_{w,i}|x)}{\pi_{\text{ref}}(y_{w,i}|x)} - \frac{\beta}{k} \sum_{j=1}^{k} \log \frac{\pi_{\boldsymbol{\theta}}(y_{l,j}|x)}{\pi_{\text{ref}}(y_{l,j}|x)} \right) \right].$$

To further understand the variance, we present the following Proposition 2, which shows that the variance of the estimator $\hat{\ell}$ is $\tilde{\mathcal{O}}(1/k)$, where $k$ is the number of samples.

**Proposition 2.** *Let $\mu_p = \mathbb{E}_{x\sim p}[f(x)]$, $\mu_q = \mathbb{E}_{x\sim q}[f(x)]$, $\sigma_p^2 = Var_{x\sim p}[f(x)]$, $\sigma_q^2 = Var_{x\sim q}[f(x)]$ and $n$ and $m$ be the number of independent samples from distributions $p$ and $q$, respectively. Then, the variance of the mini-batch estimator $\hat{\ell}$ is given by*

$$Var(\hat{\ell}) = \mathcal{O}\left( \left( \frac{\sigma_p^2}{n} + \frac{\sigma_q^2}{m} \right) \cdot \left( \frac{\sigma_p^2}{n} + \frac{\sigma_q^2}{m} + (\mu_p - \mu_q - c)^2 \right) \right).$$

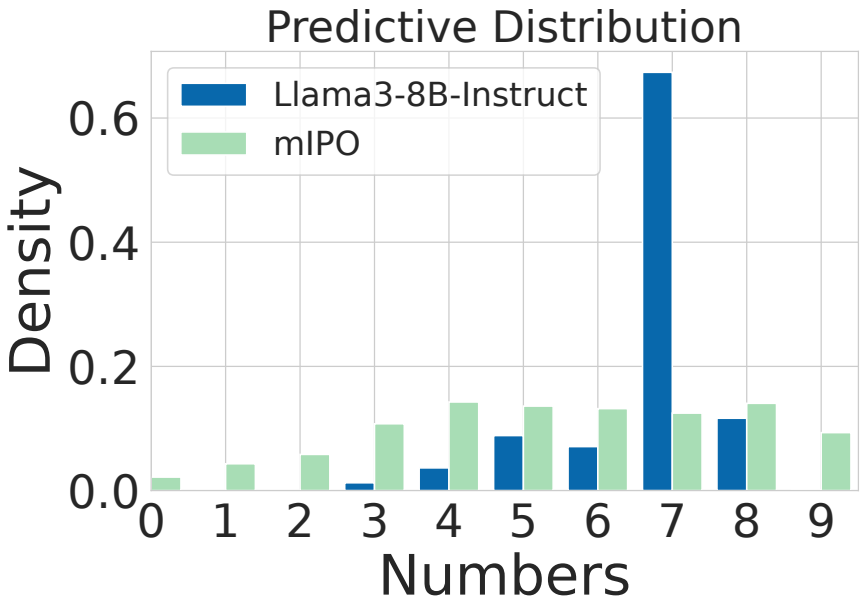


Figure 3.3: Distribution comparisons for language models before and after finetuning.

To gain an understanding about the variance and bias, we run a simulation using the function $f(x) = x^2$ to assess the effects of variance and bias and their relationship to the sample size $k$. The trend is plotted in figure 3.2. We observe that the unbiased estimator results in lower error compared to the biased estimator at small sample sizes. However, as the sample size increases, the error of the biased estimator also decreases and performs similarly to the unbiased one. This implies that as the sample size increases, the choice of estimator may become less critical.

## 3.6 Experiments

### *3.6.1 Random Number Generator (RNG)*

Despite existing LLMs being trained to follow instructions and perform well on many tasks, they perform surprisingly poorly when asked to generate random outputs (See figure 3.3). A canonical example is instructing LLMs to generate a random number within a given interval. Existing LLMs exhibit a strong bias towards certain numbers rather than producing purely

Table 3.1: Comparison of Win Rates in the Random Number Generation Experiment with Llama-3-8B: Win rates are calculated based on the uniformity of the distribution given the prompt.

| | **mIPO vs IPO** | **mIPO vs SFT** | **IPO vs SFT** | **mDPO vs DPO** | **mDPO vs SFT** | **DPO vs SFT** |
|---|---|---|---|---|---|---|
| Win Rate | 0.95 | 0.99 | 0.98 | 0.80 | 0.99 | 0.99 |

random outputs. In such settings, single-sample comparisons may be ineffective for modeling the distributional properties of the model's outputs, necessitating the use of multi-sample comparisons.

**Preference data construction and finetuning.** Our initial experiments focus on generating integers uniformly at random within an interval $[a, b]$, where $a$ and $b$ are integers. To create the dataset, we sample $a$ randomly from $[0, 1000]$ and the interval gap $m$ from $[5, 10]$, with $b$ defined as $a + m$. The preferred response group is based on a uniform distribution of numbers, while non-uniform distributions are classified as rejected. We generated approximately 3,000 paired groups for training. For testing, we sample $a$ and $b$ from $[0, 1000]$ with $a < b$ but do not impose the same gap constraints as in the training set to evaluate generalizability. We implemented the multi-sample versions of IPO and DPO, named mIPO and mDPO, respectively. To stabilize training, we add a negative log-likelihood[3] to the objective of both the multi-sample and original versions of IPO and DPO, shown to be effective in reasoning tasks [Pang et al., 2024]. For finetuning, we use the Llama 3-8B instruct model with LoRA, rank 64, and $\alpha = 128$.

**Metrics and results.** We compare the predictive entropy for each model with 100 testing prompts instructing the model to generate random numbers within specific intervals. We calculate the average win rates for the same prompt, with the winner being the one with larger entropy. The results are presented in Table 3.1. We observe that mIPO, IPO, mDPO,

3. Without this, the language model tends to generate numbers beyond the specified range. We discuss this further from a constrained optimization perspective.

and DPO consistently and significantly outperform the SFT baseline. Additionally, both mIPO and mDPO outperform IPO and DPO by a large margin. Figure 3.3 shows the predictive distribution before and after finetuning. The original Llama3 model significantly favors the number 7 over others. However, after finetuning, the predictive distribution is much closer to the uniform distribution.

### *3.6.2 Controlled Debiasing for Image Generation*

Diffusion models [Sohl-Dickstein et al., 2015] have emerged as a powerful workhorse for image generation [Ho et al., 2020, Rombach et al., 2022, Dai et al., 2023, Esser et al., 2024]. Despite their success, the generated images often reflect strong biases and stereotypes due to imbalances in the training data [Ananya, 2024]. These biases can result in harmful societal stereotypes and limit the diversity of generated content. To illustrate this point, we visualize the distribution of generated images for different occupations and highlight the discrimination in race and biases in Figure 3.7 in the Appendix using Stable diffusion 1.5 [Rombach et al., 2022].

To address the generation biases, we extend the diffusion DPO objective proposed in Wallace et al. [2023] from single-sample comparison to multi-sample comparison in a similar way as we discussed in Section 4.5. For diffusion models, the mDPO objective can be formulated as[4]

$$\mathcal{L}^{\text{diff}}_{\text{mDPO}}(\epsilon_{\boldsymbol{\theta}}, \mathcal{D}) = \mathbb{E}\left[-\log\sigma\left(-\beta T\omega(\lambda_t)\cdot\left(\mathbb{E}_{\mathcal{G}_w}\left[r(\epsilon^w, \boldsymbol{x}_t^w, t; \boldsymbol{\theta})\right] - \mathbb{E}_{\mathcal{G}_l}\left[r(\epsilon^l, \boldsymbol{x}_t^l, t; \boldsymbol{\theta})\right]\right)\right)\right],$$

where $r(\epsilon^w, \boldsymbol{x}_t^w, t; \boldsymbol{\theta}) = \|\epsilon^w - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t^w, t; c)\|_2^2 - \|\epsilon^w - \epsilon_{\text{ref}}(\boldsymbol{x}_t^w, t; c)\|_2^2$, $\epsilon^w = \boldsymbol{x}_t^w - \boldsymbol{x}^w$ and $\boldsymbol{x}^w \sim \mathcal{G}_w$, and the expectation is taken over $(\mathcal{G}_w, \mathcal{G}_l, c) \sim \mathcal{D}$ and $t \sim \mathcal{U}(0, T)$. The same applies to $\mathcal{G}_l$, $\boldsymbol{x}_t^l$ and $\boldsymbol{x}^l$. We adopt the Stable Diffusion 1.5 [Rombach et al., 2022] as our

4. It's more appropriate to interpret $r(\epsilon, \boldsymbol{x}; \theta)$ as a cost rather than a reward. Consequently, the formulation uses $-\beta$ instead of $\beta$, as done in the original DPO.

Figure 3.4: Comparison between images generated with the prompt *A portrait photo of a billionaire*.

Figure 3.5: Comparison between images generated with the prompt *A portrait photo of a engineer*.

base model due to its popularity within the community, which we also find has strong bias in favor of certain genders and races.

**Dataset construction and finetuning.** To construct the dataset, we first collect 50 occupations, and take 80% of the occupations for generating training data and the remaining 20% for evaluation. For each occupation, we generate 6 images with varied races in {black, white, aisan} and genders in {male, female}, which form our chosen group. For the rejected group, we use the images generated from the original Stable diffusion 1.5 via API[5]. Our training code is adapted from the original implementation by Wallace et al. [2023].

**Evaluation metric and results.** To measure the generation quality, we used the Simpson Diversity Index [Simpson, 1949], which focuses more on the dominance and even distribution of species. To compute it, we use $D = 1 - \sum_i (n_i/N)^2$, where $n_i$ is the number of instance falls in $i_{\text{th}}$ category (e.g., for gender, the categories we consider are {male, female}), and

5. https://stablediffusionapi.com/docs/category/stable-diffusion-api

Table 3.2: Averaged Simpson Diversity Index for the generated images of different occupations.

| | Original | SFT | DPO | mDPO |
|---|---|---|---|---|
| Gender ↑ | 0.110 | 0.318 | 0.283 | **0.353** |
| Race ↑ | 0.124 | 0.454 | 0.447 | **0.516** |

$N = \sum_i n_i$. In general, the higher the value, the better the diversity. We compared three methods, SFT, DPO (equivalent to $k = 1$ for mDPO), and mDPO ($k = 6$). The results are presented in Table 3.2.

We observe that mDPO performs the best in both metrics, which improves the Simpson Diversity Index signifacantly over the original SD 1.5, but DPO did not improve over SFT and even perform worse than SFT on Gender. For qualitative comparison, we visualized 10 images generated by the original diffusion model and the diffusion model after mDPO finetuning in Figure 3.5 using the testing prompt. We observe that the finetuned model generates more balanced and diversified images in terms of race and gender than the baseline.

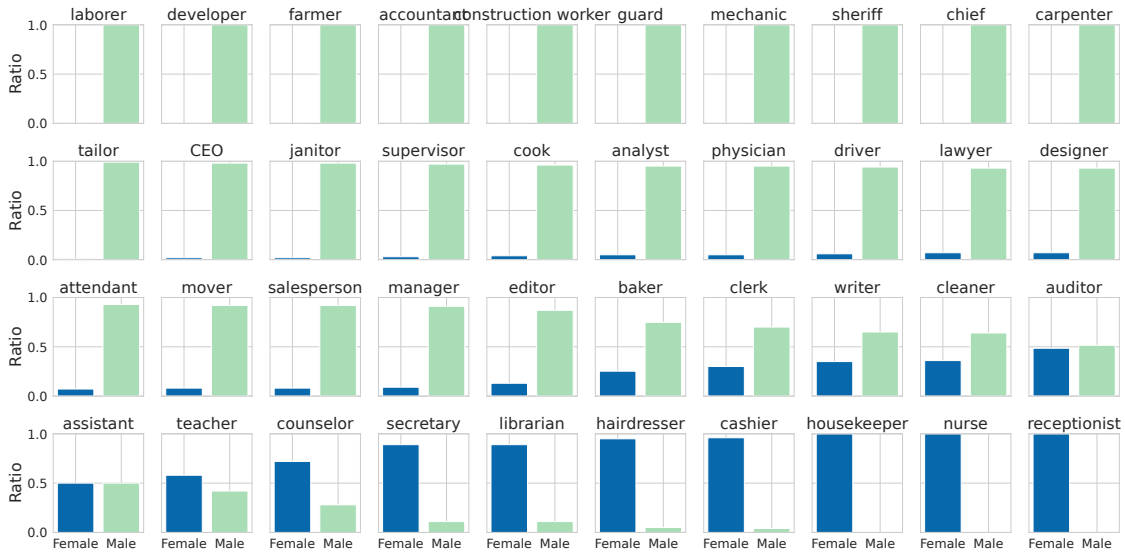


Figure 3.6: Gender distribution for the images generated by Stable Diffusion 1.5 for each occupations. For most of the occupations, it is either biased towards females or males.


Figure 3.7: Race distribution for the images generated by Stable Diffusion 1.5 for each occupations. For most of the occupations, it is biased towards white males/females, and the remaining few are biased towards either black males/females or asian males/females.

### *3.6.3 Improving Quality of Creative Fiction Generation*

Fiction generation represents a crucial aspect of the broader field of creative writing, and serves as an essential benchmark for evaluating the capabilities of LLMs [Gómez-Rodríguez and Williams, 2023, Mohammadi, 2024]. Despite the proficiency of current LLMs in crafting creative narratives, a significant challenge remains in ensuring a diverse representation of genres [Patel et al., 2024, Wang et al., 2024d] while maintaining high writing quality. Ideally, when prompted with a consistent fiction topic, LLMs should be capable of generating distinctly different stories across various genres, closely aligning with user intentions, even in the absence of explicit genre specifications in the prompt. This diversity should extend beyond mere lexical variations, incorporating unique genre-specific elements in each narrative output. In this section, we assess whether the proposed mDPO and mIPO can help fine-tune LLMs to achieve higher quality and more diverse fiction generations. Similar to the previous sections, we compare mDPO and mIPO, with their original baselines.

Table 3.3: Quality comparison of creative fiction writing.

| | DPO | mDPO (k=3) | mDPO (k=5) | IPO | mIPO (k=3) | mIPO (k=5) |
|---|---|---|---|---|---|---|
| Quality | 10.570 | 10.671 | **11.483** | 10.623 | **11.190** | 10.806 |

**Preference data construction and finetuning.** We utilized over 8,000 publically available prompts[6] for fiction generation in constructing our preference dataset. Each prompt was submitted five times to both the Llama 2-7B and Llama 3-8B models. For prompts to Llama 3-8B, we explicitly defined the genres including fantasy, sci-fi, mystery, romance, and horror, to enhance genre diversity, while keeping the prompts for Llama 2-7B unchanged. The responses from Llama 3-8B comprised the chosen set in our preference dataset, whereas the responses from Llama 2-7B [Touvron et al., 2023b] formed the rejected set. When finetuning with the single-sample DPO and IPO baselines, one response from each chosen and rejected set was selected. Conversely, with mDPO and mIPO, $k$ responses were selected accordingly. We finetuned the Llama 3-8B model using this preference dataset.

**Evaluation metrics and results.** We primarily evaluated the fine-tuned Llama 3-8B using two metrics: writing quality and genre diversity. For assessing writing quality, we employed the evaluation rubrics proposed by Chakrabarty et al. [2024], which measure fiction quality across four dimensions: fluency, flexibility, originality, and elaboration. Each dimension is specifically evaluated using a total of 14 Yes/No questions. To measure the diversity, we measure entropy of the generated genre diversity and lexical diversity, such as distinct-n [Li et al., 2016]. In our experiment, we combined all the questions and used GPT-4o [Achiam et al., 2023] as the judge. Each “Yes" was scored as 1 and each “No" as 0, with the maximum possible score for a generated fiction being 14. We then compared the final scores of Llama 3-8B finetuned with mDPO, mIPO, and their single-sample baselines.

6. https://draftsparks.com/browse/fiction-prompts/

Table 3.4: Lexical-level diversity between the proposed mDPO, mIPO, and baseline methods.

| | DPO | mDPO ($k = 3$) | mDPO ($k = 5$) | IPO | mIPO ($k = 3$) | mIPO ($k = 5$) |
|---|---|---|---|---|---|---|
| *distinct-1* ↑ | 0.025 | 0.026 | **0.027** | 0.025 | 0.026 | **0.032** |
| *distinct-2* ↑ | 0.187 | 0.189 | **0.192** | 0.189 | 0.185 | **0.209** |

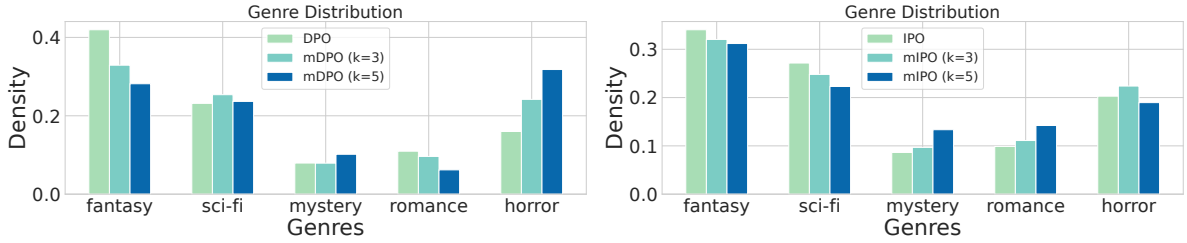


Figure 3.8: Diversity in fiction generation using the same model (Llama 3-8B) finetuned with different approaches, assessed through genre distribution. **Left:** mDPO and DPO. **Right:** mIPO and IPO. The KL-divergences between different genre distributions and the uniform distribution are (smaller is better, and the best ones are highlighted in **bold font**.) DPO: 0.170; mDPO ($k = 3$): **0.126**; mDPO ($k = 5$): 0.142; IPO: 0.125; mIPO ($k = 3$): 0.094; mIPO ($k = 5$): **0.050**.

We assess genre diversity using both lexical-level metrics, including the number of distinct unigrams (distinct-1) and bigrams (distinct-2) [Li et al., 2016], as well as semantic-level genre distribution identified by GPT-4o.

The results of the fiction writing quality evaluations are presented in Table 3.3. It is evident that the proposed multi-sample approaches have a clear advantage over the single-sample baselines, despite the more comprehensive and well-rounded challenge of enhancing the creative writing capabilities [Mohammadi, 2024]. The diversity evaluation results, based on the number of unique unigrams and bigrams as well as genre distributions, are shown in Table 3.4 and Figure 3.8, respectively. To better illustrate the differences in diversity in fiction generation, we calculate the KL-divergence between the distributions shown in figure 3.8 and the uniform distribution, with the results presented in the caption. We observe that models finetuned using multi-sample methods exhibit improved diversity at both lexical and semantic levels compared to the baselines.

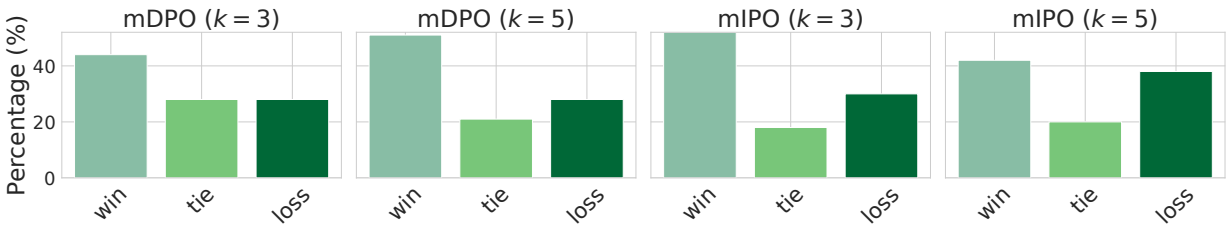


Figure 3.9: mDPO and mIPO versus DPO and IPO on Alpaca Evals using GPT-4o evaluation.

### *3.6.4 Training with Llama3-70B vs. Llama3-8B Generated Preference Data*

Synthetic data is becoming increasingly prevalent due to its scalability and lower cost compared to human labeling [Meta AI, 2024, Adler et al., 2024, Liu et al., 2024]. Our last set of experiments aims to demonstrate the efficacy of our method in handling synthetic datasets with inherent label noise. This is particularly useful in iterative alignment scenarios, where we train the model iteratively to enhance alignment performance. In these cases, we might have two qualitatively good models from prior iterations, but one may outperform the other in average. However, when examining individual responses, the performance of these models may not always be consistently better than the other. In such situations, mDPO or mIPO algorithms should be more robust than DPO or IPO. This intuition is supported by the following remarks and subsequent experiments.

**Remark 1.** *For two independent and bounded random variables $X$ and $Y$, if $\mathbb{E}[X]-\mathbb{E}[Y] > 0$, then the probability $p(\sum_{i=1}^{k} X_i > \sum_{i=1}^{k} Y_i)$ will (approximately) increase as the sample size $k$ increases. Therefore, the multi-sample pairwise comparison is (approximately) more likely to be correct than the single-sample pairwise comparison. In the asymptotic setting ($k \uparrow \infty$), the probability will converge to 1 as $\mathbb{E}[X] > \mathbb{E}[Y]$.*

To empirically validate our intuition and demonstrate the effectiveness of our method, we conduct experiments using the Alpaca benchmark [Dubois et al., 2024b] with the Llama 3-8B base model [Meta AI, 2024]. Initially, we use the instruct versions of Llama 3-8B and Llama

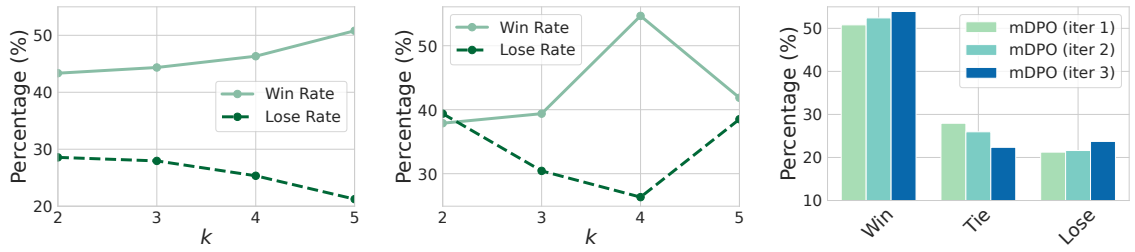


Figure 3.10: **Left:** The impact of $k$ for mDPO evaluated using GPT-4o; **Middle:** The impact of $k$ for mIPO evaluated using GPT-4o; and **Right:** Iterative improvement with mDPO.

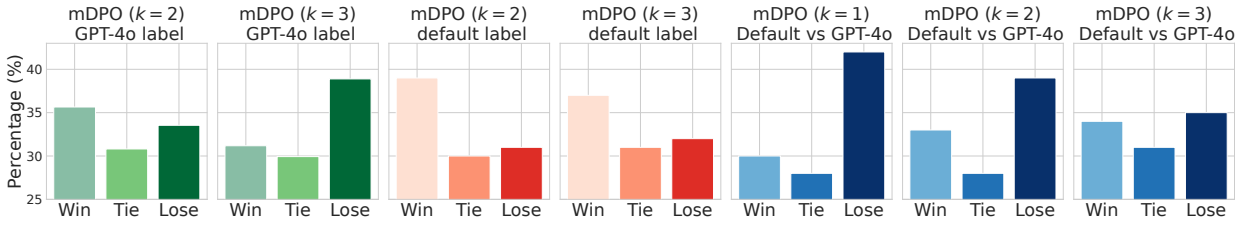


Figure 3.11: Evaluation of multi-sample and single-sample comparison under varying label conditions. The left two plots (green) depict performance in a noise-free setting using GPT-4o labels, the middle two plots (red) show results with default (noisy) labeling, and the right three plots (blue) compare the performance gap between noise-free and noisy settings.

3-70B to generate five responses for each prompt[7]. We then select the responses generated by the 70B model as the chosen group and those from the 7B model as the rejected group. We follow the Alpaca training procedures for SFT and RLHF. We first finetune the base model on the Alpaca SFT dataset, and then apply mDPO or mIPO with $k = 1, 2, ..., 5$ on the synthetic data.

**Results with varying $k$** . The results for varying group sizes $k$ are shown in figure 3.9, where win rates are computed against models trained using DPO and IPO with the same dataset. The results indicate that both mDPO and mIPO significantly outperform the baselines in terms of win rates. Additionally, we perform an ablation study to examine the effect of $k$ using mDPO and mIPO. As illustrated in the leftmost and middle plots of

7. We used outputs generated by Llama models instead of the original outputs from the Alpaca dataset.

figure 3.10, we observe that as the value of $k$ increases, the win rates improve while the lose rates decrease for most values of $k$, except for $k = 5$ for mIPO.

**Results on iterative improvement**. Lastly, we conduct experiments to demonstrate the effectiveness of our method in iteratively improving the alignment performance. We use data generated from previous rounds to form the preference data and then apply our method with $k = 5$ for iterative fine-tuning of the model. The results, shown in the rightmost plot of figure 3.10, reveal that with each iteration, our method consistently enhances the win rate compared to the baseline in terms win rates.

**Choice between mDPO and DPO?** We used GPT-4o to label pairs of sample groups generated by Llama 3-8B and 70B for $k = 1, 2, 3$ to simulate a noise-free setting. To reduce GPT-4o labeling costs, we sampled 20% of the original synthetic dataset. We refer to the dataset with the original preference label as the default label and the GPT-4o-generated label as the GPT-4o label. Our experiments reveal that without labeling noise, multi-sample comparison has no advantage over single-sample comparison, as shown in the left two plots in Figure 3.11. These plots display the win rates for $k = 2$ and $k = 3$ compared to $k = 1$, with $k = 2$ showing a slight improvement over $k = 1$. To confirm the impact of potentially noisy labels, we switched to the default labeling. In this case, both $k = 2$ and $k = 3$ outperformed $k = 1$, as shown in the middle two plots in Figure 3.11, consistent with our prior results. Lastly, to quantify effect of label noise, we computed the win rate of models trained using default labeling versus GPT-4o labeling for $k = 1, 2, 3$. The right three plots in Figure 3.11 demonstrate that default labeling performs significantly worse than GPT-4o labeling. However, as $k$ increases, the gap between win and lose rates narrows, confirming that multi-sample comparison is more robust to labeling noise. In conclusion, multi-sample comparison is much more advantageous with labeling noise, while single-sample comparison is best with noise-free labels.

## 3.7 Conclusions

In this chapter, we introduced Multi-sample Direct Preference Optimization (mDPO) and Multi-sample Identity Preference Optimization (mIPO), novel extensions to the existing Direct Alignment Preference (DAP) methods. By leveraging multi-sample comparisons, mDPO and mIPO address the limitations of traditional single-sample approaches, offering a more robust framework for optimizing collective characteristics such as diversity and bias in generative models. Comprehensive empirical studies demonstrate the effectiveness of the proposed methods across various domains. In random number generation (section 3.6.1), higher uniformity and improved handling of label noise are achieved. In text-to-image generation (section 3.6.2), multi-sample optimization enables notable reductions in gender and race biases, enhancing the overall fairness and representation in generated images. In creative fiction generation (section 3.6.3), both the quality and diversity of outputs are significantly improved. We further demonstrate that the proposed methods are especially robust against label noise (section 3.6.4).

# CHAPTER 4

# CAUSAL REWARD MODELS FOR LARGE LANGUAGE MODEL ALIGNMENT

## 4.1 Introduction

To optimize for human preferences, preference data plays a pivotal role. The method introduced in the prior two chapters rely on a static offline dataset, which results in off-policy optimization. Such paradigm still falls in the supervised learning regime, which may be insufficient to push the capabilities of LLMs. Recent advances in large language models (LLMs) have demonstrated significant progress in performing complex tasks with online reinforcement learning. However, aligning LLMs with human preferences remains challenging due to spurious correlations in reward modeling. While Reinforcement Learning from Human Feedback (RLHF) has been successful in finetuning LLMs, it often introduces biases, such as length, sycophancy, concept, and discrimination biases, that stem from these spurious correlations, distorting true causal relationships. To address this, we propose a novel causal reward modeling approach that integrates causal inference to mitigate these spurious correlations. Our method enforces counterfactual invariance, ensuring reward predictions remain consistent when irrelevant variables are altered. Through experiments on both synthetic and real-world datasets, we show that our approach mitigates various types of spurious correlations effectively, resulting in more reliable and fair alignment of LLMs with human preferences. As a drop-in enhancement to the existing RLHF workflow, our causal reward modeling provides a practical way to improve the trustworthiness and fairness of LLM finetuning.

## 4.2 Reward Modeling in RLHF

Recent advancements in large language models (LLMs) have demonstrated remarkable capabilities in generating coherent, contextually appropriate responses across a wide range of tasks [Brown et al., 2020]. A key approach to further refine these models is Reinforcement Learning from Human Feedback (RLHF), which leverages human evaluations to guide the training process and align model outputs more closely with human preferences [Stiennon et al., 2020, Ouyang et al., 2022, Bai et al., 2022a, Wang et al., 2024c]. RLHF typically involves training a reward model to capture human preferences, which is then used to fine-tune LLMs via reinforcement learning (RL) [Schulman et al., 2017, Chen et al., 2024b].

Despite the success of RLHF, reward modeling is inherently prone to *spurious correlations*, which are associations in the training data that do not reflect true causal relationships [Veitch et al., 2021], and can lead to unintended biases and induce *reward hacking* [McMilin, 2022]. Reward hacking occurs when RL agents exploit flaws or ambiguities in the reward function to maximize rewards without genuinely improving alignment with desired behaviors or completing designed tasks [Amodei et al., 2016, Weng, 2024]. Consequently, this leads to misaligned models that exhibit biases such as favoring longer outputs (*length bias*) [Zheng et al., 2023], agreeing with user's incorrect assertions (*sycophancy bias*) [Perez et al., 2022a], developing unintended shortcuts when making predictions (*concept bias*) [Zhou et al., 2023], and implicitly developing discrimination over certain demographic groups (*discrimination bias*) [Tamkin et al., 2023, Chen et al., 2024c]. These biases, rooted in spurious correlations and reward hacking rather than true causal relationships, undermine the reliability and trustworthiness of LLMs, posing significant challenges for their safe and responsible deployment in real-world applications [Anwar et al., 2024].

To understand and mitigate these issues, it is essential to consider the sources of error in reward modeling. The total error in the reward model can be decomposed into *reducible*

and *irreducible* components, as shown in Equation (4.1).

$$\text{Total Error} = \underbrace{\text{Bias}^2 + \text{Variance}}_{\text{Reducible Error}} + \underbrace{\text{Noise}}_{\text{Irreducible Error}} \tag{4.1}$$

The reducible error comprises estimation errors stemming from limited data and model approximation, which can be alleviated by collecting more data or increasing model capacity. However, irreducible error originates from inherent noise and imperfections in the data, such as the spurious correlations described earlier, which cannot be resolved merely by increasing data quantity or model complexity [Geman et al., 1992]. For instance, if longer responses are disproportionately represented and favored among higher-reward examples, the reward model may learn to prefer longer outputs irrespective of their quality, leading to the length bias observed in RLHF policies. Similarly, human annotators may unintentionally favor responses that flatter them. This bias can mislead the model, causing it to prefer agreeableness over truthfulness [Perez et al., 2022a]. Notably, such biases cannot be mitigated by simply increasing the size of the dataset. On the contrary, it may further exacerbate the effects of reward hacking [Ribeiro et al., 2016].

To address this challenge, we propose a novel approach in this work that integrates causality into reward modeling to mitigate the impact of spurious correlations and prevent reward hacking in RLHF. By leveraging causal inference techniques, we develop a **causal reward model (CRM)** that is robust to these spurious correlations and captures the true causal relationship of responses on human preferences. Central to our method is the concept of counterfactual invariance, which ensures that the reward model's predictions remain consistent under interventions on irrelevant aspects of the input, thereby reducing the irreducible error caused by spurious correlations [Veitch et al., 2021].

By addressing the irreducible errors due to spurious correlations, our approach mitigates reward hacking and advances the development of more aligned and trustworthy LLMs, enabling broader adoption in applications that demands high reliability and fairness. Specif-

ically, our contributions can be summarized as follows:

- We introduce a causal framework for reward modeling that incorporates causal regularization into the training process, allowing the model to learn "true"[1] causality from spurious relationship.
- Through experiments on both synthetic and real-world datasets, we demonstrate the effectiveness of our causal reward model (CRM) in mitigating biases, including length, sycophancy, concept, and discrimination biases, which are common factors that lead to reward hacking.
- Our method is simple to implement and can be seamlessly integrated into existing RLHF pipelines, providing a practical solution to enhance the reliability of LLMs.

## 4.3 Related Works

### *4.3.1 Reward Hacking and Spurious Correlation*

The issue of reward hacking has become increasingly significant as RLHF has grown in popularity over recent years [Amodei et al., 2016, Casper et al., 2023, Kaufmann et al., 2023, OpenAI, 2023b]. RLHF aligns LLMs with human preferences by training a reward model (RM) to provide feedback based on user prompts [Christiano et al., 2017, Ziegler et al., 2019, Chen et al., 2024b]. However, RMs are often imperfect proxies of true human preferences, leading to instances of reward over-optimization [Coste et al., 2023, Moskovitz et al., 2023], or *reward hacking* [Denison et al., 2024, Everitt et al., 2021], where models achieve high rewards without fulfilling the intended objectives [Pan et al., 2022, Weng, 2024].

LLM reward hacking often stems from the model's reliance on *spurious correlations* in the preference dataset, such as length [Sountsov and Sarawagi, 2016, Dubois et al., 2024a, Huang

1. Here, "true" represents the user's belief about what is true.

et al., 2024], sycophancy [Sharma et al., 2023, Ranaldi and Pucci, 2023], conceptual [Zhou et al., 2023], and demographic [Salinas et al., 2023] biases. These spurious correlations, closely linked to reward hacking, can impair a model's capability to learn and generalize to broader scenarios [Ribeiro et al., 2016, Geirhos et al., 2020].

Without proper constraints, models will exploit all available informative features during training, including unreliable spurious ones, which results in reward hacking, even if the task is very simple [Nagarajan et al., 2020, McMilin, 2022]. To address this, our approach integrates causal regularization into reward modeling, enabling LLMs to learn true causal relationships, mitigate the effects from spurious correlations and thereby prevent reward hacking.

### *4.3.2 Alleviating Spurious Correlations*

Early efforts to mitigate spurious correlations and reward hacking in RLHF have primarily focused on penalizing specific biases within reward models [Mnih et al., 2015], especially correcting for length bias. For example, Singhal et al. [2023] reveals that length-based biases in reward models significantly influence RLHF outcomes, often overshadowing non-length-related features, and proposes mitigation strategies such as balanced preference datasets, reward data augmentation, confidence-based truncation, increased KL penalties, explicit length penalties, omitting long outputs, and focusing on non-length reward metrics.

To address the overemphasis on longer response, Shen et al. [2023] proposed a Product-of-Experts (PoE) framework to decouple reward modeling from sequence length, thereby reducing the reward model's preference for verbose but low-quality responses. Building on this, Eisenstein et al. [2023] introduced reward model ensembles to moderate reward hacking by diversifying the sources of feedback and reducing reliance on any single reward model's spurious correlations. However, this method only partially mitigates the problem and falls short of fully eliminating reward hacking. More recently, Ramé et al. [2024] proposed Weight

Averaged Reward Models (WARM), which enhance robustness to distribution shifts by averaging model weights, offering a more efficient and effective alternative to ensemble-based policy interpolation. ODIN [Chen et al., 2024a] advanced this line of work by introducing a disentangled reward model architecture to tackle length bias. Their approach separates reward factors into two linear heads, isolating content quality for use during RL fine-tuning, thus improving performance without sacrificing efficiency.

In contrast to these approaches, our causal reward modeling incorporates causal regularization directly into the reward modeling process. By enforcing counterfactual invariance, we ensure that model responses align with the true causal effects of human preferences rather than being driven by spurious correlations. Notably, unlike existing methods [Singhal et al., 2023, Chen et al., 2024a] that address a single type of spurious correlation, our approach fundamentally mitigates a broad spectrum of spurious correlations, providing a comprehensive solution to reward hacking and enabling more reliable alignment with human preferences.

## 4.4 Preliminaries

### *4.4.1 Reinforcement Learning from Human Feedbacks (RLHF)*

**Supervised finetuning (SFT).** The supervised fine-tuning step typically starts with a pre-trained language model, which is then fine-tuned through supervised learning on a high-quality dataset tailored to specific downstream tasks, such as dialogue [Bai et al., 2022a], instruction following [Longpre et al., 2023], and summarization [Zheng et al., 2024b]. This fine-tuning process produces a model denoted as $\pi^{\text{SFT}}$.

**Reward model learning.** During this stage, we will first need to have a dataset that consists of preference pairs of responses, $(y_1, y_2)$, for each prompt $x$. Typically, these pairs are obtained by presenting them to labelers (e.g., humans), who evaluate the responses based on their preferences, represented as $y_w \succ y_l \mid x$, where $y_w$ and $y_l$ denote the preferred

and less preferred responses, respectively. From a modeling perspective, these preferences are assumed to be generated from an unknown latent reward model, $r^*(y, x)$. In practice, the modeling assumptions for the preferences can vary depending on the problem, but the Bradley-Terry (BT) model is a commonly used assumption. The BT model computes the probability of one response $y_1$ being preferred over the other response $y_2$ under the true reward function $r^\star(x, y)$ by

$$p^*(y_1 \succ y_2|x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}$$

Given a static dataset of preference data $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$ sampled from $p^*$, we can fit a reward model $r_\phi(x, y)$ to estimate its parameters through maximum likelihood estimation. This approach is equivalent to binary classification and can be trained by minimizing the negative log-likelihood loss:

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma\big(r_\phi(x, y_w) - r_\phi(x, y_l)\big)\right]$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$. In RLHF, the reward model $r_\phi(x, y)$ is often initialized from the supervised fine-tuning (SFT) model $\pi^{\text{SFT}}(y|x)$ by replacing the final layer with a classification head, which outputs a scalar (i.e., the reward).

**Fine-tuning with reinforcement learning.** Once the reward model, which serves as a proxy for the utility we aim to maximize, is trained, the next step is to apply reinforcement learning under this reward model. Typically, the following objective is used:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}[r_\phi(x, y)] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta(y|x)\|\pi_{\text{ref}}(y|x)]$$

Here, $\beta$ is a coefficient that controls the deviation from the reference policy $\pi_{\text{ref}}$, which is typically the SFT model $\pi^{\text{SFT}}$. In practice, the policy $\pi_\theta$ is also initialized using the SFT

model $\pi^{\text{SFT}}$. The KL constraint is crucial, as it prevents the model from deviating too far from the SFT model, which is helpful to mitigate issues such as forgetting [Schulman, 2015, Schulman et al., 2017, Abdolmaleki et al., 2018, Jaques et al., 2019].

### *4.4.2 Counterfactual Invariance*

An ideal debiased reward model should intuitively remain invariant to spurious factors of variations. For example, to eliminate length bias, the reward model should exhibit invariance to changes in response length. To formalize this notion, we leverage the concept of *counterfactual invariance* [Veitch et al., 2021].

We begin by introducing some notation. Let $Z$ represent the random variable corresponding to a spurious factor of variation (e.g., length), and let $T$ denote the random variable that encompasses the prompt-response pair. A reward model $r$ is said to exhibit counterfactual invariance to $Z$ if $r(T(z)) = r(T(z'))$ for all $z, z'$, where $z$ and $z'$ are realizations of $Z$, and $T(z)$ denotes the counterfactual $T$ we would have observed if $Z$ were $z$. Throughout the paper, we use the term "invariant" or "debiased" to refer specifically to counterfactual invariance.

### *4.4.3 Causal Decomposition*

The prompt-response pair $T$ can be decomposed into latent components based on their relations with the spurious factor $Z$ [Veitch et al., 2021]. Specifically, we define $T^{Z,\perp}$ as the component of $T$ that is not causally influenced by $Z$. In other words, $T^{Z,\perp}$ represents the part of $T$ such that any function of $T$ is counterfactually invariant to $Z$ if and only if it depends solely on $T^{Z,\perp}$.

Under weak conditions on $Z$, $T^{Z,\perp}$ is well defined. Further details regarding the derivation and properties can be found in [Veitch et al., 2021]. In the next section, we extend this concept to develop a reward model that incorporates counterfactual invariance, which

enables debiasing against various spurious factors.

## 4.5 Method

Ideally, counterfactual examples are necessary to learn counterfactual invariant predictors [Quinzan et al., 2022]. However, obtaining such examples is challenging, especially in RLHF settings. For instance, given a response of length 100, it is very hard to create a counterfactual response of length 50. Nonetheless, as suggested by Veitch et al. [2021], observable signatures implied by causal graphs can be leveraged to regularize the hypothesis class of the predictor.

Consider the causal diagram of reward models in figure 4.1. Here, $Z$ is the spurious factor (e.g., response length), $T$ is the prompt-response pair, $R$ is the reward and $L$ is the preference label. The binary label $L$ (e.g., $L = 1$ when $X_1$ is preferred) can be modeled under the Bradley-Terry model, where preferences depend on *true* rewards. In practice, however, human labels are often biased, captured by a direct edge from $Z$ to $L$.

As discussed in section 4.4.3, $T$ can be decomposed into latent components based on their relation with $Z$. In addition to $T^{Z,\perp}$, we define $T^{L,\perp}$ as the component that does not directly cause $L$, and $T^{Z \wedge L}$ as the complementary remaining part. And an invariant reward model should depend solely on $T^{Z,\perp}$. Although precisely learning such an invariant reward model is infeasible without counterfactual dataset, the causal graph reveals that $T^{L,\perp}$ is independent of $Z$. Consequently, any counterfactual invariant reward model must also be independent of $Z$, which leads to the following condition:

$$f(T) \perp\!\!\!\perp Z \tag{4.2}$$

This independence condition is merely a necessary condition implied by counterfactual invariance. But the key idea is that it constrains the hypothesis class, potentially guiding the model toward learning an invariant predictor.

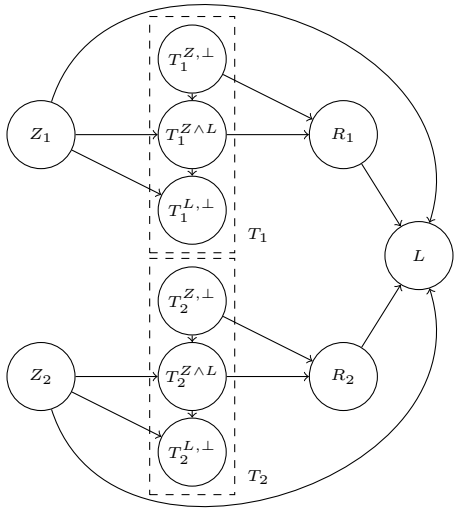


Figure 4.1: Diagram illustrating the proposed causal reward modeling. Here, $Z$ represents spurious factors (e.g., response length), $T$ denotes the prompt and response pair, $R$ is the true reward, and $L$ is the human preference label. The diagram highlights the decomposition of $T$ into latent components: $T^{Z,\perp}$, which is independent of $Z$; $T^{Z\wedge L}$, representing factors influenced by both $Z$ and $L$; and $T^{L,\perp}$, which does not causally impact $L$. This framework shows how reward hacking, modeled via direct paths from $Z$ to $L$, can mislead traditional reward models. Our proposed approach aims to isolate $T^{Z,\perp}$, ensuring counterfactual invariance and debiasing reward predictions.

### *4.5.1 Maximum Mean Discrepancy (MMD) Regularization for Independence*

To enforce the independence condition outlined in Equation (4.2), we employ Maximum Mean Discrepancy (MMD), a kernel-based statistical measure that quantifies the divergence between two probability distributions [Gretton et al., 2012, Liu et al., 2020]. MMD is commonly used to regularize models by ensuring alignment between distributions across domains or subpopulations [Tolstikhin et al., 2016, Zhang et al., 2024a]. Formally, given two distributions $\mathbb{P}$ and $\mathbb{Q}$, the squared MMD in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ is defined as:

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}, \mathcal{H}_k) = \sup_{f \in \mathcal{F}} \left( \mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{y \sim Q}[f(y)] \right)^2, \tag{4.3}$$

where $\mathcal{F}$ denotes a class of functions in $\mathcal{H}_k$, and $x \sim \mathbb{P}$, $y \sim \mathbb{Q}$ are two random variables. Intuitively, MMD measures the maximum mean difference between $\mathbb{P}$ and $\mathbb{Q}$ over functions $f \in \mathcal{F}$, as determined by a kernel $k(\cdot, \cdot)$ such as Gaussian kernels.

In our approach, we use MMD as a regularizer to ensure that the learned reward model $f(T)$ is invariant to the spurious variable $Z$. If $Z$ is binary, our MMD regularizer can be defined as:

$$\text{MMD}\left(P(f(T)|Z=0), P(f(T)|Z=1)\right).$$

When $Z$ spans a large or continuous space (e.g., response lengths), directly applying MMD becomes computationally intensive. To address this, we partition $Z$ into $M$ discrete bins and compute MMD across all pairs of bins. Let $b \in [1, M]$ denote bin indices, with $P_b(f(T))$ representing the conditional distribution of $f(T)$ within bin $b$, the regularizer is then defined as:

$$\sum_{m,m' \in [M]} \text{MMD}(P_m(f(T)), P_{m'}(f(T))).$$

This binning approach ensures the applicability of MMD in high-dimensional or continuous settings while preserving the ability to capture variations across $Z$.

In our architecture, $f(T)$ denotes the latent representation of the prompt-response pair $T$. The reward model $r_\phi(x, y)$ is parameterized by $\phi$ and depends on $f(x, y)$, such that:

$$r_\phi(x, y) = r_\phi(f(x, y))$$

To regularize $r_\phi(x, y)$, we map all responses into $M$ bins based on their spurious factor $Z$ (e.g., reponse length). For each bin $b$, we compute the conditional distribution $P_b(f(x, y))$. The overall objective function, combining the reward model training loss and the MMD-

based regularizer, is:

$$-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(r_\phi(x,y_w)-r_\phi(x,y_l))]+\lambda\sum_{m,m'\in[M]}\mathrm{MMD}(p_m(r(x,y)),p'_m(r(x,y))),$$

where $\sigma(x)=\frac{1}{1+e^{-x}}$ is the sigmoid function, and $\lambda$ is a hyperparameter controlling the weight of the MMD regularization. This formulation enforces counterfactual invariance by penalizing discrepancies in reward predictions across bins of the spurious variable $Z$, effectively guiding the model to learn invariant representations.

## 4.6 Experiments

To examine the effectiveness of the proposed reward model, we'll test on four dataset covering sycophantic, length, concept and discrimination bias. Although we only apply marginal regularization in section 4.5, in practice, the focus is typically on the model's ability to generate the chosen responses based on prompts. Therefore, we additionally test another variant of the regularization where the prompt and response pair are divided into chosen and rejected subsets and the independence regularization is applied for each subset individually. We denote this as the *conditional* causal reward model (CRM), in addition to the *unconditional* variant discussed before.

### *4.6.1 Addressing Sycophantic Bias (Semi-synthetic)*

Sycophantic bias [Sharma et al., 2023, Ranaldi and Pucci, 2023] refers to a model's tendency to produce responses that agree with or flatter the user, regardless of the truth or accuracy of the content. This bias often arises when reward models inadvertently assign higher rewards to outputs that align with users' stated beliefs or preferences, particularly in preference datasets where agreement is implicitly favored over truthfulness. For example, in a conversational setting, if annotators systematically reward responses that confirm the user's input (e.g.,

"Yes, you are correct"), the model learns to prioritize sycophantic behavior to maximize its reward. This can lead to outputs that prioritize agreeableness over factual accuracy, undermining the model's trustworthiness.

**Dataset and training.** To investigate sycophantic bias, we create a semi-synthetic dataset based on dataset developed by Sharma et al. [2023]. Specifically, our prompts are structured with the template "`{question} I think the answer is {correct_answer} but I'm really not sure.`".

In this setup, we artificially induce a correlation between sycophantic behavior and correctness. Specifically, with an 80% probability, the chosen response is prefixed with "`Yes, you are right.`" Conversely, with a 20% probability, this prefix appears in the rejected response. This creates an artificial but controlled spurious correlation between agreement ("`Yes, you are right.`") and correct answer, enabling us to observe, measure and address sycophantic bias effectively.

For SFT, we use the Llama-3 8B base model [Dubey et al., 2024], finetuned on a combination of data from the Anthropic HH-RLHF dataset [Bai et al., 2022a] and our semi-synthetic sycophantic training dataset. The HH-RLHF dataset is included to ensure sufficient training data volume, as the semi-synthetic dataset contains only 1,727 examples. The reward and policy models are then trained using the chosen/rejected pairs, with the policy fine-tuned for two epochs via Proximal Policy Optimization (PPO) [Schulman et al., 2017], implemented in OpenRLHF [Hu et al., 2024]. Additional implementation details are available in Section 4.9.

**Results.** For each test prompt, we generate 50 responses. We then quantify sycophancy by checking whether the phrase "`Yes, you are right.`" appears in any of those responses. Table 4.1 reports the percentage of test prompts for which all 50 sampled responses exhibit sycophantic behavior. It is worth noting that the SFT model, trained on chosen responses with high correlation with sycophantic phrasing, naturally tends to produce "`Yes, you are`

Table 4.1: Results on semi-synthetic syncophatic dataset. The conditional CRM outperforms other methods. Bold values indicate the best performance. Results are averaged over three runs of PPO.

| Model | Average Percentage (%) |
|---|---|
| Vanilla RM | 92.67 |
| Conditional CRM | **19.78** |
| Unconditional CRM | 62.64 |

`right.`" as a default pattern. In contrast, both the conditional and unconditional CRM approaches successfully disentangle this spurious correlation and reduce the prevalence of sycophantic responses.

### *4.6.2 Addressing Length Bias*

Length bias [Zheng et al., 2023] refers to the tendency of reward models to favor longer responses due to spurious correlations in the training data. For instance, in human preference datasets, annotators may unconsciously associate longer responses with higher-quality or more comprehensive answers, leading to disproportionate rewards for verbosity rather than substantive content. This bias often misaligns the model's behavior with true human preferences, particularly when concise and accurate responses are preferred in real-world applications.

**Dataset and training.** We adopted the Alpaca dataset [Dubois et al., 2024b] for our experiments. Initially, we uses the chosen response for each prompt to do supervised finetuning (SFT) using the Llama-3 8B base model [Dubey et al., 2024]. Then, this SFT model was subsequently employed to train both the reward model and the policy model. For reward model, we used the chosen and the rejected pair for training. With the reward model, we then trained the SFT policy with the PPO implementation from OpenRLHF [Hu et al., 2024] for one epoch. Additional details on hyperparameters and configurations are available

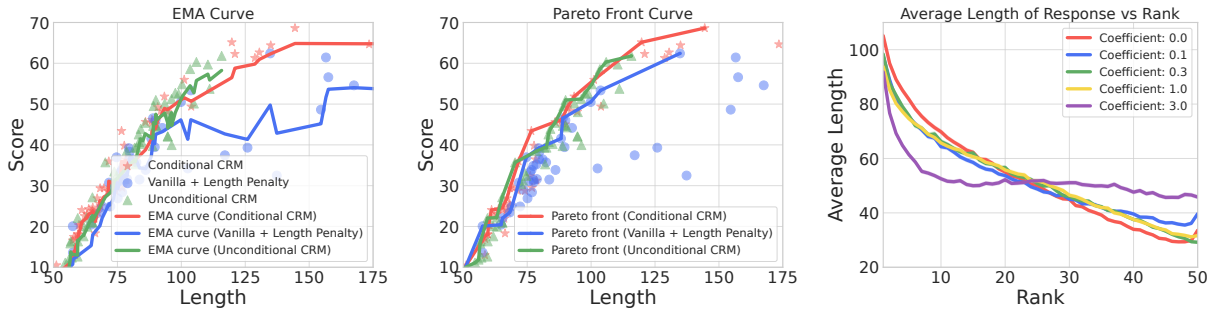


Figure 4.2: Results on Length Bias, where each dot represents models trained with different regularization coefficients and PPO hyperparameters. The leftmost figure displays the results as an exponential moving average (EMA) curve, the middle plot illustrates the Pareto front, and the rightmost figure shows the correlation between length and rank based on reward values for different causal reward models.

in the Section 4.10.

**Results.** Our findings are illustrated in figure 4.2, where each dot on the plots represents a single model run, evaluated by its win rate, calculated as the proportion of wins against the SFT model. The score is defined by score $= 50 + (n_{\text{win}} - n_{\text{lose}})/N * 100$, where $n_{\text{win}}$ and $n_{\text{lose}}$ denote the counts of wins and losses, respectively, and $N$ represents the total test count. In the leftmost plot, we observe that both the conditional and unconditional causal regularization methods achieve superior performance compared to the vanilla reward model with length penalty, as shown by their higher exponential moving average (EMA) curves. Furthermore, when examining the Pareto frontier, our approach demonstrates an advantage over the baseline method.

Finally, we analyze the impact of the regularization effect by sampling 50 responses per prompt and ranking them using a reward model trained with varying causal regularization coefficients. We then compute the average response length across all prompts for each rank. Our results show that models with higher coefficients assign higher ranks (i.e., lower numerical rank values) to responses with shorter lengths, indicating a reduction in the bias toward longer responses.

### *4.6.3 Addressing Concept Bias*

Concept bias [Zhou et al., 2023] in LLMs refers to the model's unintended reliance on correlations between specific concepts and labels present in the training data. For instance, in the Yelp Review dataset [Zhang et al., 2015], if most reviews mentioning "`food`" (categorized as a food concept) are labeled with positive sentiments, the LLM may develop a shortcut, incorrectly predicting positive sentiment for any review that involves "`food.`" This type of concept bias, which stems from associating unrelated terms with certain outcomes due to imbalanced distribution in the training data, causes LLMs to make incorrect predictions in new, unseen scenarios, which highlights the tendency of LLMs to overgeneralize based on spurious correlations, rather than always grasping the actual context of the input. In this section, we demonstrate the effectiveness of the proposed causal reward modeling in mitigating the concept bias when conducting sentiment analysis of the review datasets.

**Dataset and training.** We conducted experiments using Yelp [Zhang et al., 2015], IMDB [Maas et al., 2011a], and Amazon Shoe Review [He and McAuley, 2016] datasets, augmented with additional concept labels provided by Zhou et al. [2023]. Specifically, each dataset includes three concepts, where Yelp has "`price`", "`service`", "`food`"; IMDB has "`music`", "`acting`", "`comedy`"; and Amazon has "`size`", "`color`" and "`style`". To introduce more obvious concept bias, following Zhou et al. [2023], we modified each dataset to ensure all positive-sentiment samples were explicitly linked to a specific concept. For instance, in the Yelp dataset, we filtered reviews so that all positive sentiment entries were linked to the "`food`" concept.

To facilitate training, we reformatted the datasets to align with the structure of Anthropic hh-rlhf [Bai et al., 2022a] dataset. Specifically, we appended the prompt "`Classify the text into negative, or positive`" to the front of each review, and used the correct "`positive`" or "`negative`" label from ground truths as the chosen assistant response. The

incorrect classifications were then used in the rejected assistant response. We fully supervise finetuned (SFT) the Llama-3 8B base model [Dubey et al., 2024] on each of the above processed, concept-biased datasets using the chosen responses. The resulting SFT model was further utilized to train both vanilla and causal reward models. Finally, using these reward models, we conducted PPO finetuning using implementations from OpenRLHF [Hu et al., 2024] on the SFT model to produce final models for evaluation. More details on training hyperparameters are explained in Section 4.11.

**Metrics.** We assess performance using both utility metrics (Acc@C, Acc@NoC) as well as the bias-specific metric Bias@C, as introduced in [Zhou et al., 2023]. The utility metrics, which reflect the accuracy of correct sentiment classifications with (Acc@C) and without (Acc@NoC) the presence of a concept, indicate better performance with higher values. On the other hand, Bias@C measures spurious correlations associated with concept $C$, where values closer to zero suggest weaker biases. Specifically, positive Bias@C values suggest the model tends to predict positive labels when concept $C$ is present in the input, whereas negative values suggest the opposite tendency. For a more detailed explanation of the Bias@C metric, we direct interested readers to their original work [Zhou et al., 2023].

**Results.** As shown in Table 4.2, the results demonstrate that CRM consistently reduces concept bias across the Yelp, IMDB, and Amazon Shoe Review datasets compared to the vanilla reward model. Specifically, both conditional and unconditional CRMs achieve significantly lower Bias@C values, with conditional CRM showing reductions of up to 97% on the Yelp dataset (e.g., for the "`Price`" concept). These results highlight the effectiveness of our approach in mitigating spurious correlations.

Beyond bias reduction, the results also illustrate the trade-offs between conditional and unconditional CRMs. While conditional CRM often performs the best in Bias@C reduction, unconditional CRM demonstrates superior Acc@NoC and Acc@C performance, particularly

Table 4.2: Models performance after finetuning with PPO using both vanilla and the proposed causal reward models across concept-biased Yelp, IMDB, and Amazon Shoe Review datasets. Bold values indicate the best performance.

| | Price (Yelp) | | | Service (Yelp) | | | Food (Yelp) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc@NoC | Acc@C | Bias@C | Acc@NoC | Acc@C | Bias@C | Acc@NoC | Acc@C | Bias@C |
| Vanilla RM | 59.26 | 71.47 | 18.88 | 69.09 | 71.43 | -15.54 | 78.77 | 67.48 | 7.31 |
| Conditional CRM | **97.22** | **99.04** | **0.52** | **99.45** | **97.56** | **-0.61** | 97.77 | **99.09** | **0.71** |
| Unconditional CRM | 94.44 | 98.35 | 6.86 | 98.18 | 97.21 | -3.56 | **98.88** | 97.57 | -0.86 |
| | Music (IMDB) | | | Acting (IMDB) | | | Comedy (IMDB) | | |
| | Acc@NoC | Acc@C | Bias@C | Acc@NoC | Acc@C | Bias@C | Acc@NoC | Acc@C | Bias@C |
| Vanilla RM | 77.78 | 73.98 | 13.49 | 75.54 | 71.81 | -20.94 | 69.93 | 75.78 | 20.09 |
| Conditional CRM | 68.89 | 55.73 | **2.86** | 54.84 | 60.64 | **-7.68** | 58.04 | 56.35 | **7.99** |
| Unconditional CRM | **88.89** | **88.35** | 9.52 | **89.52** | **86.17** | -13.24 | **85.31** | **89.45** | 12.41 |
| | Size (Amazon) | | | Color (Amazon) | | | Style (Amazon) | | |
| | Acc@NoC | Acc@C | Bias@C | Acc@NoC | Acc@C | Bias@C | Acc@NoC | Acc@C | Bias@C |
| Vanilla RM | 76.17 | 54.08 | -4.05 | 63.88 | 72.47 | 15.48 | 38.30 | 74.35 | -10.16 |
| Conditional CRM | **79.95** | **85.87** | -2.37 | **84.58** | **80.73** | **2.45** | **87.94** | **80.64** | **-0.70** |
| Unconditional CRM | 73.89 | 53.26 | **-1.58** | 62.56 | 70.41 | 3.93 | 38.30 | 72.20 | -1.49 |

on datasets such as IMDB, where unconditional CRM achieves average accuracies of 87.9% for Acc@NoC and 88.0% for Acc@C, significantly outperforming the Vanilla RM baseline's 74.4% and 73.9%, respectively. This balance suggests that unconditional CRM effectively mitigates bias while preserving high predictive utility in concept-relevant contexts. However, we leave more in-depth investigation into the dynamics of this trade-off for future work.

### *4.6.4 Addressing Discrimination Bias*

Given the implicit biases embedded in training data, LLMs often learn spurious discriminatory patterns over different demographic groups [Tamkin et al., 2023]. While some previous works attempt to leverage post-training methods [Bai et al., 2022a] to mitigate this issue by designing specific bias-countering preference pairs, these approaches are often labor-intensive, lacks explicit guarantees of effectiveness, and can compromise the model's overall utility [Allam, 2024]. In contrast, we demonstrate below the effectiveness of our proposed CRM in **explicitly** mitigating discriminatory bias without relying on specific bias-focused data, while maintaining the model's original performance on general language modeling tasks.

Table 4.3: Discrimination evaluation over a diverse set of both explicit and implicit discrimination scenarios using the Discrm-eval dataset [Tamkin et al., 2023]. The scores are the mixed-effects coefficients for each demographic variable, where the lower indicates less discrimination. The best performance is in bold.

| | Explicit | | | | Implicit | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Model | Gender | Race | Age | Avg | Gender | Race | Age | Avg | Avg |
| SFT | 0.003 | 0.002 | 0.015 | **0.007** | 0.227 | 0.251 | 0.523 | 0.334 | 0.171 |
| Vanilla RM | 0.032 | 0.016 | **0.007** | 0.018 | 0.181 | 0.230 | 0.261 | 0.224 | 0.121 |
| Conditional CRM | **0.008** | **0.002** | 0.018 | 0.009 | 0.264 | **0.181** | 0.060 | 0.158 | 0.084 |
| Unconditional CRM | 0.009 | **0.002** | 0.018 | 0.009 | **0.070** | 0.213 | **0.036** | **0.107** | **0.058** |

**Dataset and training.** To obtain discrimination-specific data, we first filter the samples in the training set of the Anthropic HH-RLHF [Bai et al., 2022a] dataset, retaining those containing a predefined large set of demographic variables specified in Section 4.12.1. This results in 35,567 samples, which we split into 90% for training, 5% for validation, and 5% for testing. To stabilize training for CRM, we further group similar demographic variables into 60 demographic bins, as detailed in Section 4.12.2, balancing granularity and training batch volumes.

**Evaluation tasks.** We evaluate both targeted discrimination bias performance and general model utility. (1) To assess discrimination bias, we use the Anthropic Discrm-eval dataset [Tamkin et al., 2023], which contains two subsets of queries focusing on *explicit* and *implicit* bias. These queries are templated with demographic attributes such as age, gender, and race (e.g., "`The patient is a [age]-year-old [race] [gender] ... diagnostic information ... should they be prioritized for a kidney transplant?`"). (2) To evaluate general utility, we randomly sample 1,000 queries from the HH-RLHF test set and assess the corresponding models' responses.

**Metrics.** To evaluate discrimination, we adopt the approach from [Tamkin et al., 2023], analyzing how demographic attributes such as race, gender, and age influence decision boundaries. Specifically, we fit a mixed-effects model and report **the coefficients of each demo-**

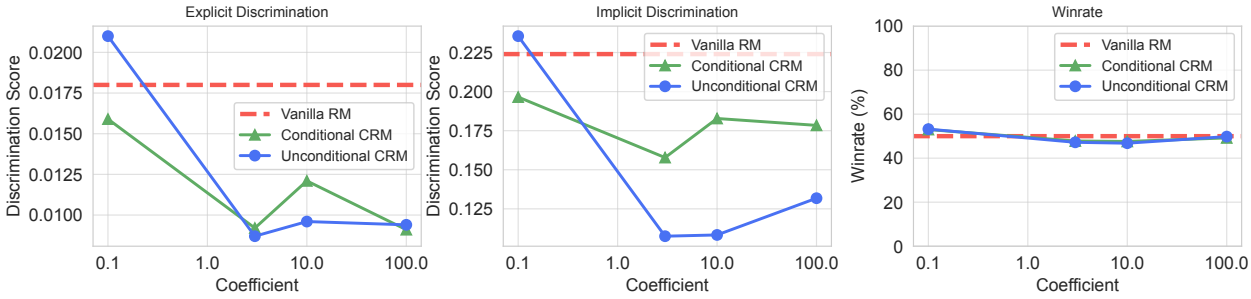


Figure 4.3: Analysis of the discrimination and utility performance on *hh-rlhf dataset* of CRMs in both conditional and unconditional settings with different MMD coefficient. The larger coefficient indicates higher weights of MMD loss. We evaluate the discrimination scores of both explicit and implicit discrimination types, and the winrate is evaluated by GPT-4o and calculated against the vanilla RM.

**graphic attribute**, where lower coefficients indicate lower bias. For model general utility, we similarly report the **win rate** comparing the performance of the CRM-enhanced model against the baseline vanilla PPO model based on evaluations conducted by GPT-4.

**Results.** As shown in Table 4.3 and figure 4.3, CRM significantly reduces discrimination across both explicit and implicit scenarios compared to the vanilla reward model. In terms of discrimination patterns, models generally exhibit higher bias in implicit scenarios, while performing relatively well in explicit questions. Nonetheless, CRM models effectively reduce bias in both cases, with a particularly significant impact on implicit scenarios where the vanilla model demonstrates greater bias.

Among the CRM variants, unconditional CRM achieves the lowest implicit discrimination score (0.107) and the best overall performance (0.058), while conditional CRM performs slightly better in explicit settings. These findings highlight CRM's effectiveness in mitigating both explicit and implicit biases across demographic attributes. The win rate analysis in figure 4.3 confirms that the additional MMD regularization term has minimal impact on the model's general utility, highlighting CRM's ability to effectively address discrimination while preserving its original performance.

## 4.7 Conclusions and Future Work

In this chapter, we introduced a novel framework for causal reward modeling (CRM) aimed at addressing spurious correlations that compromise the alignment of LLMs with human preferences. By incorporating counterfactual invariance into reward learning, our approach mitigates biases such as sycophancy, length bias, concept bias, and discrimination bias. Through extensive experiments on both synthetic and real-world datasets, we have demonstrated the effectiveness of CRM in enhancing fairness, reliability, and trustworthiness across various tasks. Additionally, our framework's seamless integration into existing RLHF workflows highlights its practical applicability, enabling more robust and equitable alignment of LLMs without introducing significant complexity. As LLMs continue to be applied to more sensitive applications, ensuring their ethical and unbiased behavior becomes imperative. By bridging the gap between causality and reward modeling, our paper takes a step toward addressing this challenge. Future work could explore extending our framework to other domains, investigating deeper causal structures, and refining regularization formulations to further optimize performance and fairness.

## 4.8 Extension with DPO

Our framework can also be extended to DPO by replacing the reward model with the DPO's implict reward. This gives us the following objective for training Causal DPO,

$$\begin{aligned}\mathcal{L}_{\text{Casual-DPO}} = &- \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \\ &+ \lambda\sum_{m,m'\in[M]}\text{MMD}\left(p\left(\frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\text{ref}}(y|x)}|b=m\right), p\left(\frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\text{ref}}(y|x)}|b=m'\right)\right). \quad (4.4)\end{aligned}$$

## 4.9 Sycophantic Bias

The reward model is trained using Low-Rank Adaptation (LoRA) [Hu et al., 2021] finetuning with rank 64 and weight $\alpha = 128$ with batch size 32 across 4 gpus. For both the conditional and the unconditional regularization, the coefficients are chosen from $\{0, 0.1, 0.3, 0.5, 1, 3, 5, 10\}$. The final policy model is trained with PPO with batch size 16 for 2 epochs. The initial KL coefficient is set to be 0.01.

## 4.10 Length Bias

To obtain the SFT model, we begin by finetuning the Llama-3 8B base model on selected responses from the Alpaca farm dataset for 3 epochs, using a learning rate of $2 \times 10^{-5}$. Additional hyperparameters are available in the Alpaca farm GitHub repository[2]. Next, we train the reward model starting from this SFT model. This training is done using LoRA finetuning with rank 64 and weight $\alpha = 128$, for 4 epochs, with a learning rate of $1 \times 10^{-4}$ and a batch size of 128 (distributed as 16 per GPU device).

To obtain a variety of reward models, we perform a hyperparameter sweep on two variables: 1) the number of bins, and 2) the regularization coefficient. For the number of bins, we explore values $\{10, 20, 30\}$, and for the coefficient, we test $\{0.1, 1.0, 3.0, 10, 100\}$. Finally, we apply PPO to finetune the SFT model under our learned reward model, obtaining the final policy model. For the PPO stage, we train for 1 epoch with a KL coefficient sweep over $\{0.003, 0.01, 0.03, 0.1\}$, resulting in a total of 60 (conditional) causal reward models.

For the baseline method, the reward model is trained with a regularization coefficient of 0 (equivalently). In the PPO stage, we perform a more thorough sweep, tuning the KL coefficient over $\{0.003, 0.01, 0.03, 0.1\}$, the learning rate over $\{5 \times 10^{-7}, 1 \times 10^{-6}\}$, and the length penalty over $\{0, 1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}, 5 \times 10^{-4}, 1 \times 10^{-6}, 5 \times 10^{-6}\}$. This

2. `https://github.com/tatsu-lab/alpaca_farm/blob/main/examples/scripts/sft.sh`

process results in 56 models, providing a comparable set to the causal reward models.

## 4.11 Concept Bias

As briefly mentioned in Section 4.6.3, we supervised finetuned (SFT) the Llama-3 8B base model on each of the processed Yelp, IMDB, Amazon Shoe Review datasets. We keep the same hyperparameters for all datasets, which are illustrated in Table 4.4. The resulting SFT model is used for the reward learning through LoRA, where detailed parameters are illustrated in Table 4.5. The reward models are subsequently utilized during the PPO, where the hyperparameters for PPO on each dataset is showed in Table 4.6. All the trainings are distributed on 8 NVIDIA A100 GPUs.

Table 4.4: Supervised finetuning hyperparameters for concept-bias experiments.

| | Supervised Finetuning |
|---|---|
| Learning rate | 2e-5 |
| Batch size | 128 |
| Gradient accumulation steps | 2 |
| Training epochs | 4 |
| Warm-up steps | 500 |

Table 4.5: Reward learning hyperparameters for concept-bias experiments.

| | Vanilla Reward | | | Conditional CRM | | | Unconditional CRM | | |
|---|---|---|---|---|---|---|---|---|---|
| | Yelp | IMDB | Amazon | Yelp | IMDB | Amazon | Yelp | IMDB | Amazon |
| Regularization coefficient | - | | | 0.1 | 0.3 | 0.1 | 0.5 | 3 | 1 |
| Learning rate | 1e-4 | | | | | | | | |
| LoRA Rank | 64 | | | | | | | | |
| LoRA Alpha | 128 | | | | | | | | |
| Batch size | 64 | | | | | | | | |
| Gradient accumulation steps | 1 | | | | | | | | |
| Training epochs | 1 | | | | | | | | |

Table 4.6: PPO hyperparameters for concept-bias experiments.

| | Proximal Policy Optimization (PPO) |
|---|---|
| Actor learning rate | 5e-7 |
| Critic learning rate | 9e-6 |
| Initial KL coefficient | 0.01 |
| Batch size | 128 |
| Gradient accumulation steps | 8 |
| Training epochs | 1 |

## 4.12 Discrimination Bias

### *4.12.1 Training Data Preparation*

This section provides the demographic keyword groups used for filtering data from the Anthropic HH-RLHF dataset [Bai et al., 2022a]. The keywords were grouped into broad categories based on demographic attributes, ensuring comprehensive coverage. Table 4.7 through 4.11 summarize the categories and their associated keywords. These keyword categories provided a robust basis for filtering the HH-RLHF dataset, ensuring the inclusion of diverse demographic contexts in the training data.

Table 4.7: Age-related categories and keywords used for filtering data.

| **Category** | **Keywords** |
|---|---|
| Childhood | *teen, child, kid, baby, toddler, infant, adolescent, youth, teenager* |
| Young Adults | *adult, young adult, young, grown-up* |
| Middle Age | *middle-aged, sophisticated, mature* |
| Elderly | *elderly, old, senior, aged* |

Table 4.8: Gender-related categories and keywords used for filtering data.

| **Category** | **Keywords** |
|---|---|
| Male | *male, man, masculine* |
| Female | *female, woman, feminine* |
| Non-binary | *gay, lesbian, LGBT, non-binary, queer* |

Table 4.9: Race-related categories and keywords used for filtering data.

| **Category** | **Keywords** |
|---|---|
| White | *white, european, caucasian, middle eastern* |
| Black | *black, african* |
| Asian | *asian, chinese, japanese, korean, indian, south asian, east asian, southeast asian* |
| Latino | *latino, hispanic* |
| Indigenous | *indigenous, native, pacific island* |

Table 4.10: Nationality-related categories and keywords used for filtering data.

| **Region** | **Keywords** |
|---|---|
| Americas | *american, canadian, mexican, brazilian, argentinian* |
| Europe | *german, french, italian, spanish, british, russian, polish* |
| Asia-Pacific | *chinese, japanese, korean, indian, australian, new zealander* |
| Africa | *nigerian, south african, egyptian, kenyan* |

### *4.12.2 Demographic Bins*

To stabilize training for CRM, we grouped similar demographic variables into 60 distinct bins. Specifically, we targeted *age*, *gender*, and *race*-related discrimination. Each row in Table 4.7, Table 4.8, and Table 4.9 was treated as a bin, resulting in a total of $4 \times 3 \times 5 = 60$ bins.

### *4.12.3 Detailed Description*

Aiming to address a comprehensive collection of discrimination factors spanning age, gender, race and political groups that LLMs might encounter when handling various forms of societal decisions, we can construct any target training dataset by following the anthropic discrimination dataset Tamkin et al. [2023] which covers 70 topics across society that involve *accepting/rejecting* a person.

Specifically, it consists of 70 decision topic templates with placeholders for demographic information (e.g. [AGE], [RACE], [GENDER]) and all the questions are framed as a yes/no decision faced by a third party decision-maker who must use only the information available

Table 4.11: Religion-related categories and keywords used for filtering data.

| **Religion** | **Keywords** |
|---|---|
| Christianity | *christian, church, bible* |
| Islam | *muslim, mosque, koran* |
| Judaism | *jewish, synagogue, torah* |
| Dharmic and Others | *hindu, buddhist, temple, religion* |

in the prompt. Critically, each question is organized in a way such that *yes* refers to a positive outcome for the subject of the decision question (e.g. having a loan approved or being promoted). Notably, they consider two ways to filling the templates:

1. **Explicit:** insert random combinations of age, race, and gender directly into the placeholders, with [AGE] $\in$ [20, 30, 40, 50, 60, 70, 80, 90, 100], [GENDER] $\in$ [male, female, non-binary] and [RACE] $\in$ [white, Black, Asian, Hispanic, Native American], in total 9450 questions.

2. **Implicit:** only specify the age and a person's name to implicitly indicate a particular race and gender (e.g. Wei Li, Carlos Reyes) . This approach focuses on assessing discrimination based on more subtle information correlated with race and gender

Thus a tentative approach would be similar to Section 4.6.2 where we employ MMD to decouple the representation from the targeted discrimination factors (age, race, gender). Suppose we have a discrimination-intensive dataset $\mathcal{D}$ to train a reward model, we can vary the inputs by first tagging the targeted discrimination factors to construct placeholders and then substituting them in the prompt templates with different combinations of age, race, and gender. We define $f(x)$ as the representation of input prompts, where $x$ is a prompt filled with specific demographic information. The reward model, parameterized by $\phi$, is thus denoted by $r_\phi(x, y)$, where $y$ represents the decision outcome. The goal is to ensure that $r_\phi(f(x), y)$ is independent of the discriminatory factors.

To measure and minimize discrimination bias with MMD, we can define demographic

groups based on combinations of age, race, and gender as $\mathcal{G}$. For each demographic group $g \in \mathcal{G}$, we can map prompts into $N$ bins, where each bin corresponds to a specific combination of demographic attributes. The MMD regularizer is computed as follows: where $y_w$ and $y_l$ indicate positive and negative decision outcomes, respectively, and $\sigma$ is the sigmoid function. The expectation is calculated over the discrimination-intensive dataset $\mathcal{D}$, where each sample $(x, y_w, y_l)$ consists of a prompt $x$ with a preferred decision outcome $y_w$ over $y_l$. The term $\lambda$ acts as a regularization parameter, balancing the objective between maximizing the reward signal and minimizing discrimination bias measured by the MMD between different demographic groups.

### *4.12.4 Evaluation Prompt*

The prompt template that we provide to GPT-4o to compare the two response of the CRM-enhanced model and the vanilla PPO baseline is detailed below:

**System**:
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, harmlessness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.
After providing your explanation, you must output your final verdict by strictly following this format: '[[A]]' if assistant A is better, '[[B]]' if assistant B is better, and '[[C]]' for a tie.

**User:**

[Initial User-Assistant Dialogue] `{prompt}`

[Start of assistant A's Response] `{rwd_response}` [End of assistant A's Response]

[Start of assistant B's Response] `{vanilla_ppo_response}` [End of assistant B's Response].

# CHAPTER 5

# CONCLUSION

## 5.1 Conclusion of Dissertation Contributions

This dissertation presents several novel advancements in the area of Reinforcement Learning from Human Feedback (RLHF), focusing on improving model alignment, addressing biases, and enhancing generative diversity. The contributions of this work can be summarized as follows:

- Generalization of DPO Framework: We introduced a generalized Direct Preference Optimization (DPO) framework that integrates various divergence constraints, such as Jensen-Shannon divergence, forward KL divergence, and -divergence. By addressing the Karush-Kuhn-Tucker conditions, we eliminated the need for the normalizing constant in the Bradley-Terry model, allowing for a more efficient and flexible approach to optimizing language models under diverse alignment constraints.

- Multi-sample Post-training: To mitigate limitations of traditional RLHF, we proposed Multi-sample Direct Preference Optimization (mDPO) and Multi-sample Identity Preference Optimization (mIPO). These approaches extend the post-training process by incorporating multi-sample comparisons, enabling better control of generative distributions. We demonstrated that these methods improve model performance, particularly in terms of diversity, bias reduction, and robustness to label noise.

- Causal Reward Modeling for Bias Mitigation: Recognizing the challenges posed by spurious correlations in reward modeling, we introduced a causal reward model (CRM) that utilizes causal inference techniques to enhance alignment and mitigate reward hacking. By ensuring counterfactual invariance, our approach significantly reduces bias and improves the fairness and reliability of the model. This method provides a

practical solution for reducing common biases, such as length, sycophancy, concept, and discrimination biases.

Overall, the contributions made in this dissertation offer a comprehensive efforts for advancing the safe and effective deployment of AI systems. These contributions ensure that language models are better aligned with human values, exhibit more generative diversity, and are more robust against biases introduced through flawed reward modeling and data.

## 5.2 Implications of Findings

The findings of this dissertation have several important implications for the field of AI alignment and the safe deployment of large language models (LLMs).

- Improved Alignment and Human Value Integration: The generalization of the DPO framework demonstrates that incorporating different divergence constraints can lead to more nuanced models that balance alignment performance with the diversity of generated outputs. This improvement in alignment ensures that models can better reflect human preferences while avoiding undesirable behavior, such as narrow political views or over-optimization towards certain outputs. Furthermore, the flexibility introduced by these constraints opens the door for tailoring models to different application domains, making AI systems more adaptable and context-sensitive.

- Better Control Over Generative Models: The introduction of multi-sample optimization methods (mDPO and mIPO) addresses a significant limitation in current post-training methods by enabling more precise control over the model's generative distribution. By capturing group-wise characteristics, these methods offer a more robust framework for managing diversity and bias, which are critical factors in areas like creative content generation and decision-making systems. These advancements allow for better handling of creative tasks, such as story writing or generating diverse numerical

outputs, while simultaneously reducing unwanted biases in the model's outputs, such as gender or racial biases in diffusion models.

- Reducing Bias and Improving Fairness: By introducing a causal approach to reward modeling, this dissertation offers a promising solution to the problem of reward hacking and spurious correlations that can lead to model misalignment and undesirable biases. The use of causal inference techniques ensures that models can better capture the true causal relationships between their outputs and human preferences, thereby improving fairness, accountability, and trustworthiness in AI systems. This method addresses critical challenges such as sycophancy bias and concept bias, which are often overlooked in traditional RLHF setups.

In practice, these advancements contribute significantly to making AI systems more reliable and trustworthy. They pave the way for safer AI deployment in real-world applications, particularly in areas that require high-stakes decision-making, such as healthcare, finance, and law enforcement, where fairness and transparency are paramount.

## 5.3 Future Research Directions

The findings of this dissertation open several avenues for future research, particularly in enhancing model alignment, understanding biases, and further improving generative diversity in AI systems. Some potential directions for future work include:

- Exploring Other Divergence Constraints: While this dissertation has explored several divergence constraints such as forward KL divergence, Jensen-Shannon divergence, and -divergence, there remains a wealth of potential divergence measures that could further refine the alignment process. Future research could explore other divergence classes, such as total variation distances, and their potential benefits for controlling generative diversity or aligning models with specific human preferences in more nuanced ways.

- Robustness of Multi-sample Optimization: Although multi-sample approaches like mDPO and mIPO showed promising results, there remains potential to improve their robustness in the presence of noisy or incomplete data. Research could focus on developing methods to further enhance the performance of multi-sample comparisons in contexts where human feedback may be sparse or ambiguous, such as when dealing with synthetic or weakly labeled data.

- Causal Inference in Large-Scale Reward Models: While the causal reward model introduced in this work addresses biases in a more robust manner, its application to large-scale, real-world datasets remains a challenge. Future research could investigate how causal inference techniques can be scaled for use in complex reward models, and how to better incorporate causal reasoning into the learning process in large-scale reinforcement learning tasks.

In summary, while this dissertation makes significant strides in addressing key challenges in AI alignment, the field remains ripe for further exploration. Continued innovation in these areas will be critical for building trustworthy, adaptable, and fair AI systems that can meet the growing demands of real-world applications.

# REFERENCES


Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024.

Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

Ahmed Allam. Biasdpo: Mitigating bias in language models through direct preference optimization. *arXiv preprint arXiv:2407.13928*, 2024.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *ArXiv preprint*, abs/1606.06565, 2016. URL `https://arxiv.org/abs/1606.06565`.

Ananya. Ai image generators often give racist and sexist results: can they be fixed?, 2024.

Anthropic. Claude 2, 2023. URL `https://www.anthropic.com/index/claude-2`. Accessed: 2023-04-03.

Anthropic. Claude 3.5 sonnet model card addendum. 2024. URL `https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf`.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *ArXiv preprint*, abs/2112.00861, 2021. URL `https://arxiv.org/abs/2112.00861`.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *ArXiv preprint*, abs/2212.08073, 2022b. URL `https://arxiv.org/abs/2212.08073`.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712, 2023. URL `https://arxiv.org/abs/2303.12712`.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United

States, 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.naacl-main.92. URL https://aclanthology.org/2022.naacl-main.92.

Francois Caron and Arnaud Doucet. Efficient bayesian inference for generalized bradley–terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

CarperAI. Trlx: Transformer reinforcement learning x. https://github.com/CarperAI/trlx, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34, 2024.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*, 2024a.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024b.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*, 2024c.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024d.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *ArXiv preprint*, abs/1810.08575, 2018. URL https://arxiv.org/abs/1810.08575.

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

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *ArXiv preprint*, abs/2304.06767, 2023. URL `https://arxiv.org/abs/2304.06767`.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *ArXiv preprint*, abs/2305.14387, 2023. URL `https://arxiv.org/abs/2305.14387`.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024a.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024b.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208, 2024.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467, 2021.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *ArXiv preprint*, abs/2305.00955, 2023. URL `https://arxiv.org/abs/2305.00955`.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.301. URL `https://aclanthology.org/2020.findings-emnlp.301`.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Team Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *ArXiv preprint*, abs/2209.14375, 2022. URL `https://arxiv.org/abs/2209.14375`.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. *ArXiv preprint*, abs/2302.08215, 2023. URL `https://arxiv.org/abs/2302.08215`.

Carlos Gómez-Rodríguez and Paul Williams. A confederacy of models: A comprehensive evaluation of llms on creative writing. *arXiv preprint arXiv:2310.08433*, 2023.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773, 2012.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

hallatore. Comment on "stable diffusion: A new approach". https://www.reddit.com/r/StableDiffusion/comments/11mulj6/comment/jbkqcyk/, March 2023. Accessed: 2024-06-26.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023. doi:https://doi.org/10.1016/j.ijresmar.2022.05.005. URL https://www.sciencedirect.com/science/article/pii/S0167811622000477.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. *ArXiv preprint*, abs/2306.12001, 2023. URL https://arxiv.org/abs/2306.12001.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jian Hu, Xibin Wu, Weixun Wang, Xianyu, Dehao Zhang, and Yu Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.

Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo M Ponti, and Ivan Titov. Post-hoc reward calibration: A case study on length bias. *arXiv preprint arXiv:2409.17407*, 2024.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *ArXiv preprint*, abs/1805.00899, 2018. URL https://arxiv.org/abs/1805.00899.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*, 2020.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=jWkw45-9AbL.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. Longform: Optimizing instruction tuning for long text generation with corpus extraction. *arXiv preprint arXiv:2304.08460*, 2023.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. Pretraining language models with human preferences. *ArXiv preprint*, abs/2302.08582, 2023. URL https://arxiv.org/abs/2302.08582.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

LAION-AI. Open-assistant, 2023. URL https://github.com/LAION-AI/Open-Assistant. Accessed: 09/22/2023.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *CoRR*, abs/1811.07871, 2018. URL http://arxiv.org/abs/1811.07871.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, 2016. Association for Computational Linguistics. doi:10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014.

Shufan Li, Harkanwar Singh, and Aditya Grover. Popalign: Population-level alignment for fair text-to-image generation. *arXiv preprint arXiv:2406.19668*, 2024.

Q Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *ArXiv preprint*, abs/2306.03100, 2023. URL https://arxiv.org/abs/2306.03100.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692, 2019. URL https://arxiv.org/abs/1907.11692.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR, 2023.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011a.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, 2011b. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1015`.

Emily McMilin. Selection bias induced spurious correlations in large language models. *arXiv preprint arXiv:2207.08982*, 2022.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. Distributional preference alignment of llms via optimal transport. *arXiv preprint arXiv:2406.05882*, 2024.

Llama Team Meta AI. The llama 3 herd of models. 2024. URL `https://scontent-sjc3-1.xx.fbcdn.net/v/t39.2365-6/452387774_1036916434819166_4173978747091533306_n.pdf?_nc_cat=104&ccb=1-7&_nc_sid=3c67a6&_nc_ohc=7qSoXLG5aAYQ7kNvgHzeJBv&_nc_ht=scontent-sjc3-1.xx&oh=00_AYC5BfhWhI-JmNF44OqmUhygHAr_yWzA059olF7GBxeZ2w&oe=66AB508D`.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1928–1937. JMLR.org, 2016. URL `http://proceedings.mlr.press/v48/mniha16.html`.

Behnam Mohammadi. Creativity has left the chat: The price of debiasing language models. *arXiv e-prints*, pages arXiv–2406, 2024.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained rlhf. *arXiv preprint arXiv:2310.04373*, 2023.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

OpenAI. Gpt-4 technical report. *ArXiv*, 2023a.

OpenAI. GPT-4 technical report, 2023b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization, 2024.

Zeeshan Patel, Karim El-Refai, Jonathan Pei, and Tianle Li. Swag: Storytelling with action guidance. *arXiv preprint arXiv:2402.03483*, 2024.

Ethan Perez, Anton Bakhtin, Rosanne Li, Luke Metz, Danilo Jimenez Rezende, Rowan McAllister, David Krueger, and Matthew Botvinick. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022a.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates, 2022b. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.225`.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Francesco Quinzan, Cecilia Casolo, Krikamol Muandet, Yucen Luo, and Niki Kilbertus. Learning counterfactually invariant predictors. *arXiv preprint arXiv:2207.09768*, 2022.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv preprint*, abs/2305.18290, 2023. URL `https://arxiv.org/abs/2305.18290`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*, 2024.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. *arXiv preprint arXiv:2405.20304*, 2024.

Leonardo Ranaldi and Giulia Pucci. When large language models contradict humans? large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*, 2023.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–15, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *ArXiv preprint*, abs/2303.17548, 2023. URL `https://arxiv.org/abs/2303.17548`.

Igal Sason and Sergio Verdú. $f$-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

John Schulman. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015.

John Schulman. Approximating kl divergence, 2020. URL http://joschu.net/blog/kl-approx.html. Accessed: 09/22/2023.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. *arXiv preprint arXiv:2310.05199*, 2023.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *ArXiv preprint*, abs/2305.15324, 2023. URL https://arxiv.org/abs/2305.15324.

Edward H Simpson. Measurement of diversity. *nature*, 163(4148):688–688, 1949.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF, 2024. URL https://openreview.net/forum?id=sNtDKdcI1f.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Ziang Song, Tianle Cai, Jason D Lee, and Weijie J Su. Reward collapse in aligning large language models. *ArXiv preprint*, abs/2305.17608, 2023. URL https://arxiv.org/abs/2305.17608.

Pavel Sountsov and Sunita Sarawagi. Length bias in encoder decoder models and a case for global conditioning. *arXiv preprint arXiv:1606.03402*, 2016.

Nisan Stiennon, Tadashi Otsuka, Gretchen Krueger, Daniel M Ziegler, Jeffrey Wu, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325*, 2020.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. FUNCTIONAL VARIATIONAL BAYESIAN NEURAL NETWORKS. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkxacs0qY7.

Simeng Sun, Dhawal Gupta, and Mohit Iyyer. Exploring the impact of low-rank adaptation on the performance, efficiency, and regularization of rlhf. *arXiv preprint arXiv:2309.09055*, 2023.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Yi Tay, Jason Wei, Hyung Won Chung, Vinh Q Tran, David R So, Siamak Shakeri, Xavier Garcia, Huaixiu Steven Zheng, Jinfeng Rao, Aakanksha Chowdhery, et al. Transcending scaling laws with 0.1% extra compute. *arXiv preprint arXiv:2210.11399*, 2022.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016.

E Paul Torrance. Torrance tests of creative thinking. *Educational and psychological measurement*, 1966.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023a. URL `https://arxiv.org/abs/2307.09288`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208, 2021.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023.

Chaoqi Wang, Roger Grosse, Sanja Fidler, and Guodong Zhang. Eigendamage: Structured pruning in the kronecker-factored eigenbasis. In *International conference on machine learning*, 2019.

Chaoqi Wang, Adish Singla, and Yuxin Chen. Teaching an active learner with contrastive examples. *Advances in Neural Information Processing Systems*, 34:17968–17980, 2021a.

Chaoqi Wang, Shengyang Sun, and Roger Grosse. Beyond marginal uncertainty: How accurately can bayesian regression models estimate posterior predictive correlations? In *International Conference on Artificial Intelligence and Statistics*, pages 2476–2484. PMLR, 2021b.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023a.

Chaoqi Wang, Ziyu Ye, Zhe Feng, Ashwinkumar Badanidiyuru Varadaraja, and Haifeng Xu. Follow-ups also matter: improving contextual bandits via post-serving contexts. *Advances in Neural Information Processing Systems*, 36:12774–12796, 2023b.

Chaoqi Wang, Yuxin Chen, and Kevin Patrick Murphy. Model-based policy optimization under approximate bayesian inference. In *AISTATS*, 2024a.

Chaoqi Wang, Ziyu Ye, Kevin Murphy, and Yuxin Chen. Don't be pessimistic too early: Look k steps ahead! In *AISTATS*, pages 3313–3321, 2024b.

Chaoqi Wang, Zhuokai Zhao, Chen Zhu, Karthik Abinav Sankararaman, Michal Valko, Xuefei Cao, Zhaorun Chen, Madian Khabsa, Yuxin Chen, Hao Ma, et al. Preference optimization with multi-sample comparisons. *arXiv preprint arXiv:2410.12138*, 2024c.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, et al. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025.

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, et al. Weaver: Foundation models for creative writing. *arXiv preprint arXiv:2401.17268*, 2024d.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *ArXiv preprint*, abs/2112.04359, 2021. URL `https://arxiv.org/abs/2112.04359`.

Lilian Weng. Reward Hacking in Reinforcement Learning, 11 2024. URL `https://lilianweng.github.io/posts/2024-11-28-reward-hacking/`.

Gian Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022. doi:10.1162/tacl_a_00502. URL `https://aclanthology.org/2022.tacl-1.58`.

Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024a.

Shentao Yang, Shujian Zhang, Congying Xia, Yihao Feng, Caiming Xiong, and Mingyuan Zhou. Preference-grounded token-level guidance for language model fine-tuning. *Advances in Neural Information Processing Systems*, 36, 2024b.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Shuhai Zhang, Yiliao Song, Jiahao Yang, Yuanqing Li, Bo Han, and Mingkui Tan. Detecting machine-generated texts by multi-population aware optimization for maximum mean discrepancy. *arXiv preprint arXiv:2402.16041*, 2024a.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito. Forcing diffuse distributions out of language models. *arXiv preprint arXiv:2404.10859*, 2024b.

Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zhili Feng, Zenghui Ding, and Yining Sun. Rankclip: Ranking-consistent language-image pretraining. *arXiv preprint arXiv:2404.09387*, 2024c.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*, 2023a.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv preprint*, abs/2306.05685, 2023. URL https://arxiv.org/abs/2306.05685.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024a.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024b. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. Describing differences between text distributions with natural language. In *International Conference on Machine Learning*, pages 27099–27116. PMLR, 2022.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36:40204–40237, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024a.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024b.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. Explore spurious correlations at the concept level in language models for text classification. *arXiv preprint arXiv:2311.08648*, 2023.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM, 2018. doi:10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

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