
A Comprehensive Survey of Direct Preference Optimization: Datasets, Theories, Variants, and Applications

Wenyi Xiao^{1*}, Zechuan Wang^{1*}, Leilei Gan^{1*†}, Shuai Zhao², Zongyue Li¹, Ruirui Lei¹, Wanggui He³, Luu Anh Tuan², Long Chen³, Hao Jiang³, Zhou Zhao¹, Fei Wu¹

¹Zhejiang University, China;

²Nanyang Technological University, Singapore; ³Alibaba Group, China;

ABSTRACT

With the rapid advancement of large language models (LLMs), aligning policy models with human preferences has become increasingly critical. Direct Preference Optimization (DPO) has emerged as a promising approach for alignment, acting as an RL-free alternative to Reinforcement Learning from Human Feedback (RLHF). Despite DPO's various advancements and inherent limitations, an in-depth review of these aspects is currently lacking in the literature. In this work, we present a comprehensive review of the challenges and opportunities in DPO, covering theoretical analyses, variants, relevant preference datasets, and applications. Specifically, we categorize recent studies on DPO based on key research questions to provide a thorough understanding of DPO's current landscape. Additionally, we propose several future research directions to offer insights on model alignment for the research community. An updated collection of relevant papers can be found on <https://github.com/Mr-Loevan/DPO-Survey>.

1 Introduction

Through pre-training on extensive, high-quality corpora using the next token prediction objective with a huge amount of computational costs, Large Language Models (LLMs) OpenAI [2022], Touvron et al. [2023a], OpenAI [2024a], Jiang et al. [2023] assimilate comprehensive world knowledge into their internal parameters, demonstrating impressive language understanding and generation abilities. Further, LLMs have been extended to accommodate multi-modality inputs, including both language and vision, thereby giving rise to Large Vision Language Models (LVLMs) OpenAI [2023a], Liu et al. [2024a], Team et al. [2023], Bai et al. [2023]. These foundation models serve as a versatile solution and have achieved exceptional performance across a broad spectrum of both language and visio-language tasks, marking a significant milestone in the advancement toward artificial general intelligence.

As these foundation models grow larger and more powerful, they are still found grappling with following the user's instruction (explicit objective) and fulfilling the goals of being Helpful, Honest and Harmless (implicit objective), which attribute to the *misaligned* next token prediction task used in the pre-training stage Leike et al. [2018], Askell et al. [2021], OpenAI [2023b]. Therefore, a typical post-training stage, known as preference optimization (e.g., Reinforcement Learning from Human Feedback, RLHF), is additionally performed at the response level to align pre-trained language models with the user's intentions and ensure they remain helpful, honest, and harmless Ouyang et al. [2022a], Dai et al. [2024], Sun et al. [2023]. RLHF first trains an explicit reward model on collected human preferences. Subsequently, RLHF fine-tunes the policy model (i.e., the LLM targeted for fine-tuning) with reinforcement learning (RL) algorithms (e.g., Proximal Policy Optimization(PPO; Schulman et al. [2017a])) to output responses which can maximize the response reward rated by the reward model but not deviate too far from a reference model constrained by KL-divergence. Nevertheless, RLHF requires meticulous hyper-parameters tuning and extensive computational resource to maintain the RL training stability. Moreover, some research has identified several challenges associated with this explicit reward modeling, such as *reward hacking* Casper et al. [2023], *reward misspecification* Pan et al. [2022] and *out-of-distribution generalization* Tien et al. [2023].

*Equal contribution.

†Corresponding author.

To avoid the aforementioned limitations of RLHF, a range of RL-free preference optimization methods have been proposed. Yuan et al. [2023], Dong et al. [2023], Liu et al. [2024b], Song et al. [2024] propose sampling multiple responses from the policy model and rating them with a well-trained reward model. Then, instead of using an RL algorithm, the policy model is directly fine-tuned under the supervision of the best-of- K response (known as reject sampling) or by applying a ranking loss to the ranked responses. On the other hand, starting from the KL-constrained reward maximization objective in RL, Direct Preference Optimization (DPO; Rafailov et al. [2023]) derive its learning objective, specifically a simple maximum likelihood objective, on offline preference data, which is directly formulated over the policy model and a reference model, thereby bypassing the explicit reward model training phase and eliminating the need for reinforcement learning optimization. Indeed, the optimization objective of DPO is equivalent to the Bradley Terry model Bradley and Terry [1952a] with an implicit reward function that is parameterized by the policy model itself. Compared to RLHF, DPO has been demonstrated stable, performant, and computationally lightweight in various applications Rafailov et al. [2023], Ethayarajh et al. [2024], Ivison et al. [2024].

More recently, some studies have indicated that despite avoiding computationally expensive reinforcement learning, DPO still encounters some substantial challenges. For instance, implicit reward modeling in DPO might lead to a biased policy that favors out-of-distribution responses Xu et al. [2024a], Saeidi et al. [2024], offline DPO is empirically inferior to online alignment methods Ivison et al. [2024], models that undergo alignment might experience an alignment tax Lin et al. [2024a], Lu et al. [2024a], et. etc. Consequently, various improved versions of DPO have been proposed recently, including KTO Ethayarajh et al. [2024], IPO Azar et al. [2023], CPO Xu et al. [2024b], ORPO Hong et al. [2024], simPO Meng et al. [2024], and others Lu et al. [2024b], Xiao et al. [2024], Zeng et al. [2024]. With the rapid progress in DPO, there is an urgent need for a comprehensive review to help researchers identify emerging trends and challenges in this field. We have observed several concurrent studies on LLM alignment that are relevant to our work Ji et al. [2023], Wang et al. [2023a], Shen et al. [2023]. However, the existing review papers primarily focus on the overall alignment of LLMs, including instruction fine-tuning and RLHF. The sections of these studies related to DPO are insufficient to capture the rapid advancements currently unfolding in this area. Furthermore, these reviews tend to focus on alignment within the context of language models, without providing a thorough introduction to the applications and datasets specific to DPO.

To bridge this gap, in this work, we present a comprehensive review of recent advancements in DPO, covering relevant preference datasets, theoretical analyses, variants, and applications. Specifically, we categorize current studies on DPO based on the following research questions:

- **Effect of Implicit Reward Modeling.** DPO circumvents the need to train an explicit reward model by establishing a direct mapping from reward functions to optimal policies. Consequently, studies have examined the generalization capabilities of the implicit reward modeling employed in DPO Lin et al. [2024b], Li et al. [2024a], Yang et al. [2024a], Jia [2024].
- **Effect of KL Penalty Coefficient and Reference Model.** Both the optimization objective of the RL and DPO involve a Kullback-Leibler (KL) divergence regularization, which constrains the policy model to remain within a specified proximity to the reference model. Therefore, some recent studies have investigated the impact of the KL penalty coefficient and the choice of the reference model Liu et al. [2024c], Xu et al. [2024a], Feng et al. [2024], Rafailov et al. [2024a].
- **Effect of Different Feedback.** DPO uses point-wise reward and pair-wise preference data to provide reward signals. However, obtaining high-quality pair-wise preference data is both costly and time-consuming, posing challenges for scalability. Additionally, instance-level optimization may not fully leverage the potential of preference data. Therefore, some studies employ other forms of feedback (e.g. List-wise, Binary, Step-wise, Token-wise, etc) as the reward signal for optimization Dong et al. [2023], Yuan et al. [2023], Ethayarajh et al. [2024], Zeng et al. [2024], Chen et al. [2024a], Xu et al. [2024b].
- **Online DPO.** Compared to online RLHF, DPO utilizes pre-collected preference data and is considered an offline preference optimization method. Some studies have highlighted the performance gap between online and offline algorithms Tang et al. [2024], Wang et al. [2024a]. To address this gap, recent research has explored iterative and online variants of DPO, as well as strategies for efficiently collecting new preference datasets Xu et al. [2024c], Guo et al. [2024a], Yuan et al. [2024a], Chen et al. [2024b].
- **Reward Hacking.** Reward hacking is a long-standing problem in RL where the policy achieves a high reward but fails to meet the actual objective Dubois et al. [2024], Singhal et al. [2023]. Recent studies have found reward hacking also exists in both RLHF and DPO regardless the explicit or implicit reward model, which exploits potential shortcuts (e.g., response length and style) to develop specific response patterns to hack the reward model Kabir et al. [2024], Wang et al. [2023b], Park et al. [2024]. To overcome this limitation, some methods have been proposed to avoid such weakness been utilized Park et al. [2024], Yuan et al. [2024b], Meng et al. [2024], Liu et al. [2024d].

- **Alignment Tax.** Preference optimization aims to align models with human preferences. However, prior studies have found a phenomenon known as the alignment tax, which refers to that improvements in alignment objective can lead to a decrease in performance compared to a SFT model Ouyang et al. [2022a]. Consequently, some studies have investigated the alignment tax and proposed methods to reduce its effect Lin et al. [2024a], Lou et al. [2024a], Guo et al. [2024b].

We hope our review will help researchers capture new trends and challenges in this field, explore the potential of DPO in aligning LLMs and Multi-modal LLMs (MLLM), and contribute to building a more scalable and generalizable DPO. Specifically, we believe that future research should prioritize the development of more advanced DPO variants that: (i) go beyond instance-level feedback to capture more fine-grained and accurate rewards; (ii) exhibit competitive or superior generalization capabilities compared to online RLHF by leveraging data, learning objectives, and rewards; and (iii) facilitate the development of sophisticated applications, such as deep reasoning systems like OpenAI o1 OpenAI [2024b], mixed-modal models like Chameleon Team [2024].

The rest of the paper is organized as follows. (§ 2) provides the background of RLHF and DPO. In (§ 3), we introduce the research questions and different variants on DPO. The used datasets and applications of DPO are presented in (§ 4) and (§ 5), respectively. (§ 6) provides the discussion on the opportunities and challenges of DPO. Finally, a brief conclusion is drawn in (§ 7).

2 Preliminary

In this section, we will revisit the foundational concepts of **RLHF** Ouyang et al. [2022a] (Reinforcement Learning from Human Feedback) and **DPO** Rafailov et al. [2023] (Direct Preference Optimization) to highlight the necessity of DPO fine-tuning as a solution to RLHF's shortcomings.

2.1 Reinforcement Learning from Human Feedback

The typical RLHF pipeline for LLMs generally consists of three key phases: (1) supervised fine-tuning (SFT), (2) preference sampling and reward model training, and (3) reinforcement learning (RL) optimization.

In the SFT stage, RLHF typically starts with a pre-trained language model, which is further refined using supervised fine-tuning on a curated dataset of high-quality human-generated responses(labels). This supervised fine-tuning stage produces an initial model, denoted as π^{SFT} , which has improved but still not fully aligned with human preferences. This model serves as the baseline for the subsequent RLHF stages. The preference learning stage involves collecting human feedback in the form of preference data. Given pairs of responses generated by π^{SFT} , human evaluators express a preference for one response over the other. These preferences serve as the foundational feedback for reward model training. The reward model's objective is to quantify these preferences by assigning a numerical score to each output, effectively converting human feedback into a scalar reward.

One widely used formula in this stage is the Bradley-Terry Bradley and Terry [1952b] (BT), which proposes a mechanism for modeling pairwise comparisons and assigning concrete reward values to the outputs. In the context of RLHF, BT is used to predict human preferences between pairs of responses. The resulting reward function is then used to guide the concrete model's decision-making by providing feedback on how well each generated response aligns with human preferences. The BT formula proposes that the human preference distribution p^* for a given response pair can be expressed as the following formula:

$$p^*(y_1 \succ y_2 | x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} \quad (1)$$

Reward Modelling. Given the dataset of prompts and pairs of $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$, a reward model $r_\phi(y, x)$ can be parameterized to predict the alignment of a given response y with human preferences. The parameters of the reward model can be optimized using the maximum likelihood estimation. The loss function used for this estimation can be expressed as the following formula:

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))] \quad (2)$$

In EQ.(2), σ represents the sigmoid function, which transforms the difference between the reward values of the winning and losing responses into a probability that reflects the model's confidence in the preference. The difference in rewards, $r_\phi(x, y_w) - r_\phi(x, y_l)$, is passed through the sigmoid function to map this difference to a probability in the range [0,1], corresponding to the model's confidence that y_w is the preferred response. The negative sign in the loss function ensures that minimizing the loss corresponds to maximizing the likelihood of correctly predicting the human preference. This

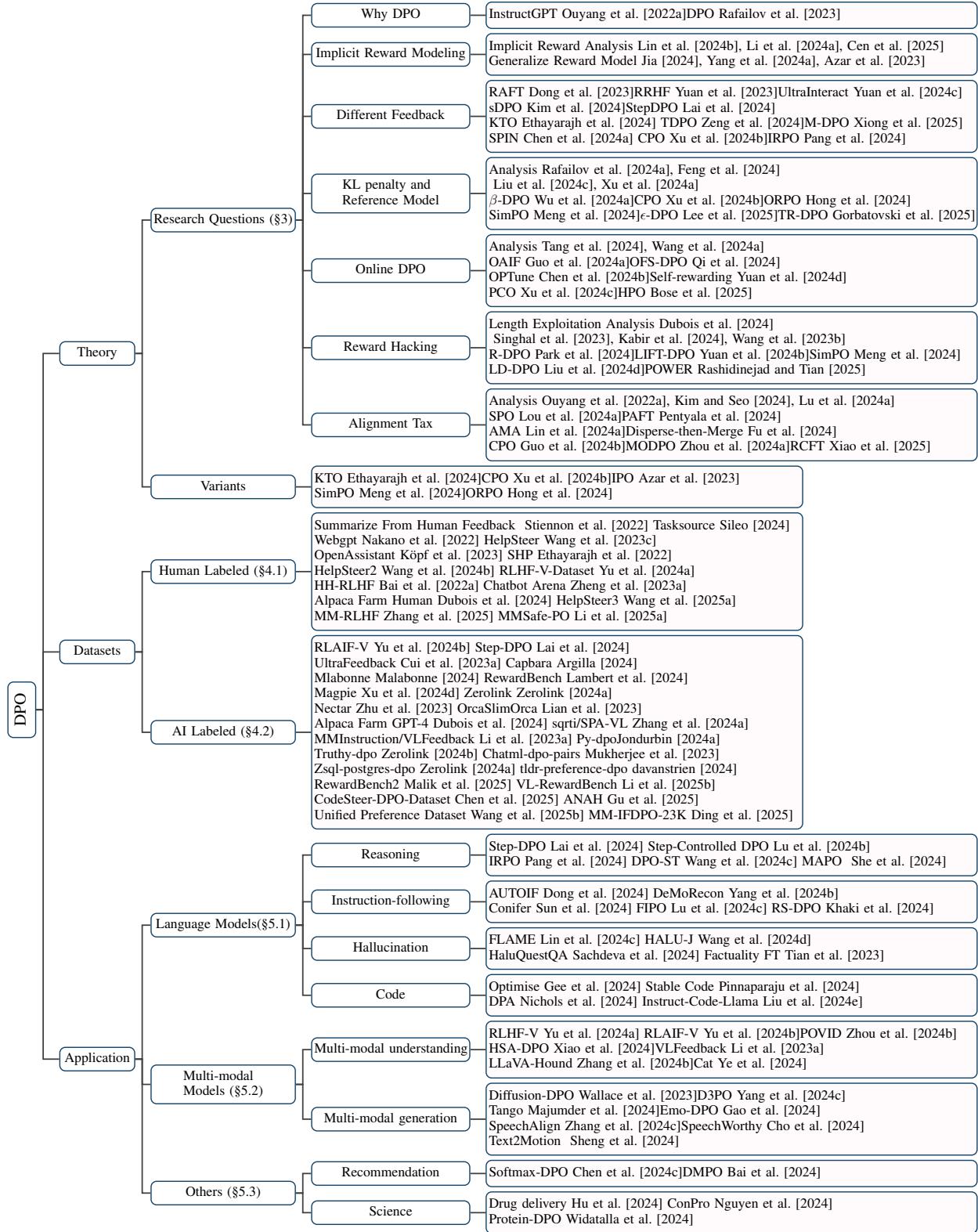


Figure 1: Taxonomy of research in DPO that consists of theory analysis, variants, datasets and applications

loss function is central to the reward model's training, guiding it to assign higher rewards to responses that are more aligned with human preferences.

RLHF Objective. Next, the learned reward model is utilized to provide feedback during the RL fine-tuning phase. The primary optimization objective in this phase is to enhance the performance of the language model based on the reward model’s feedback. The objective function can be formulated as follows:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi_\theta(y|x) \| \pi_{\text{ref}}(y|x)] \quad (3)$$

In this formulation:

- π_θ represents the policy of the language model, parameterized by θ .
- $\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)]$ denotes the expected reward, where $r_\phi(x, y)$ is the reward model that quantifies the alignment of the response y with human preferences given the prompt x .
- π_{ref} denotes the reference model, which serves as a baseline or starting point for the optimization.
- $\mathbb{D}_{\text{KL}} [\pi_\theta(y|x) \| \pi_{\text{ref}}(y|x)]$ is the Kullback-Leibler (KL) divergence between the current policy π_θ and the reference model π_{ref} .
- β is a hyperparameter that balances the reward maximization with the KL divergence penalty.

The reward model $r_\phi(x, y)$ is derived from the training process discussed previously, where it learns to predict human preferences based on a dataset of paired responses. The KL divergence term $\mathbb{D}_{\text{KL}} [\pi_\theta(y|x) \| \pi_{\text{ref}}(y|x)]$ acts as a regularization penalty, preventing the model from deviating excessively from the reference model. Without this constraint, the language model might focus solely on generating high-reward responses, which could potentially lead to outputs that are high-scoring but not necessarily useful or diverse. The goal of the RLHF objective is, therefore, twofold: to maximize the reward signal derived from the reward model and to ensure that the language model does not diverge too drastically from the reference model. This dual focus helps achieve a balance between generating responses that are both high in quality and aligned with human preferences, while also preserving the model’s foundational characteristics.

2.2 Direct Preference Optimization

In RLHF, the process is relatively intricate, involving the training of a reward model and the iterative sampling from the language model’s policy during the training loop. This complexity arises from the need to continuously evaluate and refine the model based on feedback from the reward model, resulting in significant computational demands. Direct Preference Optimization (DPO) offers a streamlined alternative by optimizing the same objective as RLHF but bypasses the explicit need for a separate reward model, thereby reducing the computational costs associated with aligning the LLM.

DPO Objective. Deriving from the KL-constrained reward maximization objective in EQ.(3), DPO has the mathematically equivalent form as the following equation:

$$\pi_r(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right) \quad (4)$$

where $Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$ is the partition function. The partition function $Z(x)$ normalizes the policy distribution $\pi_r(y|x)$. It is calculated by summing over the exponential terms of the reward function weighted by the reference model’s distribution for all possible responses y .

This summation ensures that $\pi_r(y|x)$ is a valid probability distribution, but it is impractical to enumerate all possible responses to traverse all possible input x , which requires significant computational resources, especially when handling a large response space. Actually, DPO algorithm mitigates the need for explicit reward model training and sampling from the LM policy. By directly optimizing preferences without the intermediate step of training a reward model, DPO simplifies and reduces the overall computational burden, making it a more efficient approach for language model alignment. To be specific, the objective function in (4) can be rearranged to isolate $r(x, y)$, yielding:

$$r(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \quad (5)$$

In this form, $r(x, y)$ is expressed in terms of the optimized policy π_r , the reference model π_{ref} , and the partition function $Z(x)$. By substituting this reparameterization into the original reward function r^* , we obtain:

$$r^*(x, y) = \beta \log \frac{\pi_r^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \quad (6)$$

Then the original loss of reward function r^* of will be substituted for the new reparameterization in Eq.(6). The preference model Bradley-Terry model only depends on the difference of rewards between two completions and thus the human preference probability will have the following form:

$$p^*(y_1 \succ y_2 | x) = \frac{1}{1 + \exp \left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} \right)} \quad (7)$$

This can be simplified using the sigmoid function, i.e., $p^*(y_1 \succ y_2 | x) = \sigma(r^*(x, y_1) - r^*(x, y_2))$. Here, the partition function $Z(x)$ is effectively canceled out in the subtraction, leaving a straightforward difference in reward values. Following this simplification, a maximum likelihood objective for DPO can be formulated. The resulting policy objective or loss function for DPO is:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right] \quad (8)$$

In this formation:

- π_θ denotes the policy of the language model being optimized.
- π_{ref} represents the reference model's policy.
- y_w and y_l are the winning and losing responses, respectively.

The DPO loss function is derived directly from the reward differences and avoids the need for a separate reward model. By using this approach, the policy π_θ can be analytically optimized based on the reward function differences, simplifying the optimization process and eliminating the need for an intermediate reward model. This makes DPO computationally more efficient, as it directly leverages preference data to guide the model's learning without the added complexity of reward model training.

3 Research Questions and Variants

In this section, we discuss key research questions on DPO to provide a thorough understanding of its current landscape. We start by comparing DPO with RLHF to highlight the advantages and limitations of DPO. We then investigate the effects of implicit versus explicit reward modeling, focusing particularly on generalization challenges. Additionally, we discuss the effects of different feedback and analyze the roles of the KL penalty coefficient and reference model. Lastly, we review advancements in Online DPO and discuss issues like reward hacking and alignment tax.

RQ0: why DPO? Large Language Models have recently developed rapidly with the supervised fine-tuning (SFT) training Chung et al. [2022]. However, due to the next token prediction *misaligned* with the objective *"follow the user's instructions helpfully and safely"* Brown et al. [2020], Fedus et al. [2022], SFT models often exhibit unintended behaviors. For example, they might generate toxic, harmful, hallucinatory, and biased responses Chung et al. [2022]. Furthermore, SFT solely instructs models to learn from correct responses while neglecting to address incorrect ones, thus lacking the consideration of human values or preferences Zhao et al. [2023], Ren and Sutherland [2025].

A potential solution to this issue is Reinforcement Learning from Human Feedback (RLHF), using human preferences as a reward signal to fine-tune models by reinforcement learning Chung et al. [2022], Bai et al. [2022b], Lee et al. [2023], such as PPO Schulman et al. [2017b] and actor-critic Mnih et al. [2016], Haarnoja et al. [2018]. We follow the mainstream work Ziegler et al. [2019], Stiennon et al. [2020] and focus on PPO Schulman et al. [2017b] in this paper.

Nevertheless, RLHF requires loading four models (policy, reference, reward, and critic) and involves training sensitive hyperparameters along with substantial and costly human oversight to develop a well-trained reward model. Furthermore, some research has identified several challenges associated with reward modeling. For instance, *reward misspecification* Pan et al. [2022] occurs when a scalar score from the reward model fails to comprehensively represent human preferences. Additionally, *misgeneralization* Tien et al. [2023] arises when reward models compute rewards using unexpected or contingent features of the environment Michaud et al. [2020], leading to causal confusion and poor out-of-distribution generalization. Another issue is *reward hacking* Casper et al. [2023], for example, without regularization to penalize the KL divergence between a base model and the fine-tuned model, large language models (LLMs) undergoing RL often learn to output nonsensical text Stiennon et al. [2020].

A powerful alternative that does not require an explicit reward model or RL to RLHF is Direct Preference Optimization (DPO) Rafailov et al. [2023]. DPO derived the closed-form solution of the PPO optimization objective, revealing the relationship between the reward and the optimal policy model. It then reparameterizes the reward modeling loss with

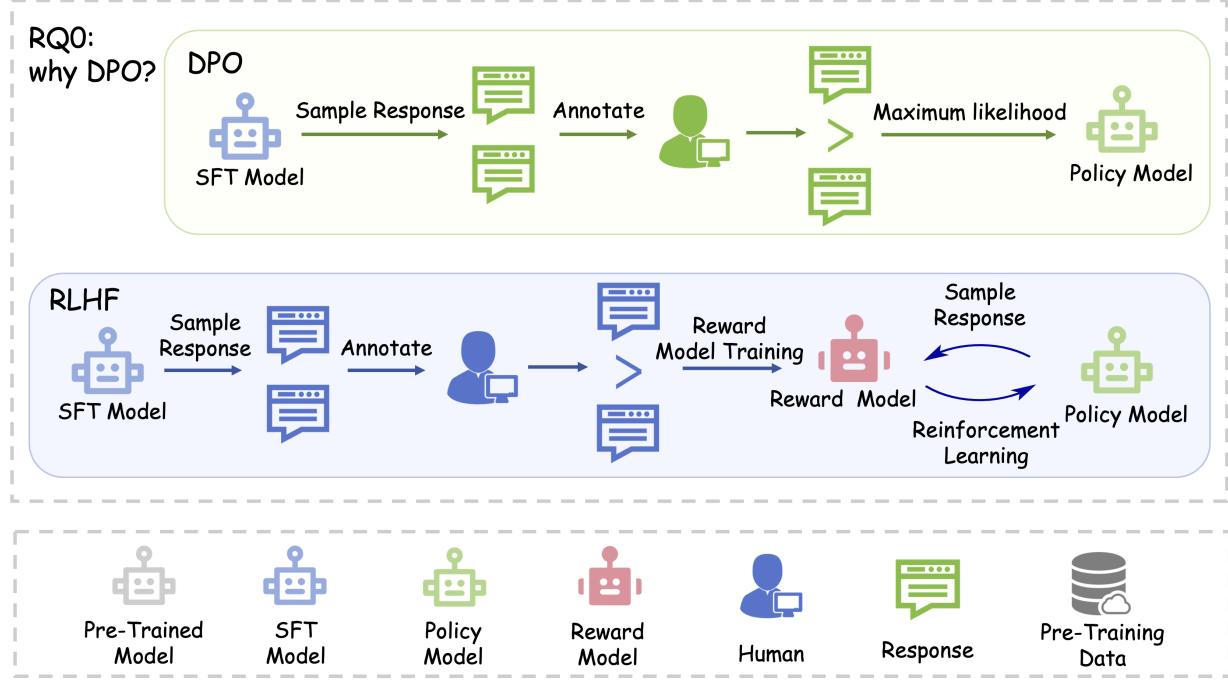


Figure 2: RQ0 Why DPO? The figure shows the pipeline of RLHF and DPO. DPO derives a closed-form solution for the optimal policy under the RLHF objective, which allows it to reparameterize the reward function in terms of the policy itself, thereby converting RLHF’s multi-stage process of explicit reward modeling and reinforcement learning into a single-stage direct policy optimization on preference data. The symbols used in this figure are consistent across all subsequent figures in this section.

the policy, thereby calculating implicit rewards and optimizing the policy model π_θ over the preference data as depicted in EQ.8.

Compared to RLHF, DPO is empirically stable, performant, and computationally lightweight Rafailov et al. [2023], Ethayarajh et al. [2024]. However, recent studies indicate that, despite avoiding explicit reward modeling, DPO faces challenges similar to those encountered in RLHF. For example, the offline preferences can lead DPO to overfit the training data distribution, thereby limiting its generalization capabilities Azar et al. [2023]. DPO is sensitive to the distribution shift between the base model outputs and preference data. Besides, DPO is prone to generating a biased policy that favors out-of-distribution responses, leading to unpredictable behaviors, etc. Xu et al. [2024a], Saeidi et al. [2024].

RQ1: Effect of Implicit Reward Modeling. We first discuss the generalization ability of the implicit reward modeling in DPO Wang et al. [2025c]. DPO avoids training an explicit reward model by establishing a mapping from reward functions to optimal policies. This enables us to transform a loss function over reward functions into a loss function over policies. $r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ is the reward implicitly defined by the policy model π_θ and reference model π_{ref} .

However, Lin et al. [2024b] find that even though implicit reward modeling in DPO (DPORM) fits the training dataset comparably, it generalizes less effectively than explicit reward modeling (EXRM), especially when the validation datasets contain distribution shifts. Their experimental results showed that across five out-of-distribution settings, DPORM showcased an average accuracy drop of 3% and a maximum drop of 7%.

IPO Azar et al. [2023] points out the overfitting reward problem in DPO and proposes a novel loss to this issue. Li et al. [2024a] conducted a theoretic and empirical analysis of the errors inherent in policy optimization methods when learning from user preferences for alignment in reinforcement learning including PPO, DPO, and IPO. Their experimental results underscore the critical importance of optimizing policies using out-of-distribution preference data. Furthermore, the study demonstrates the effectiveness of employing an explicit reward model to enhance policy performance.

Jia [2024] proposed to optimize a general Reward Modeling (RM) through a meta-learning approach. A bilevel optimization algorithm is utilized during meta-training to learn an RM that guides policy learning to align with human preferences across various distributions. Yang et al. [2024a] proposed a novel approach to enhance the reward model’s

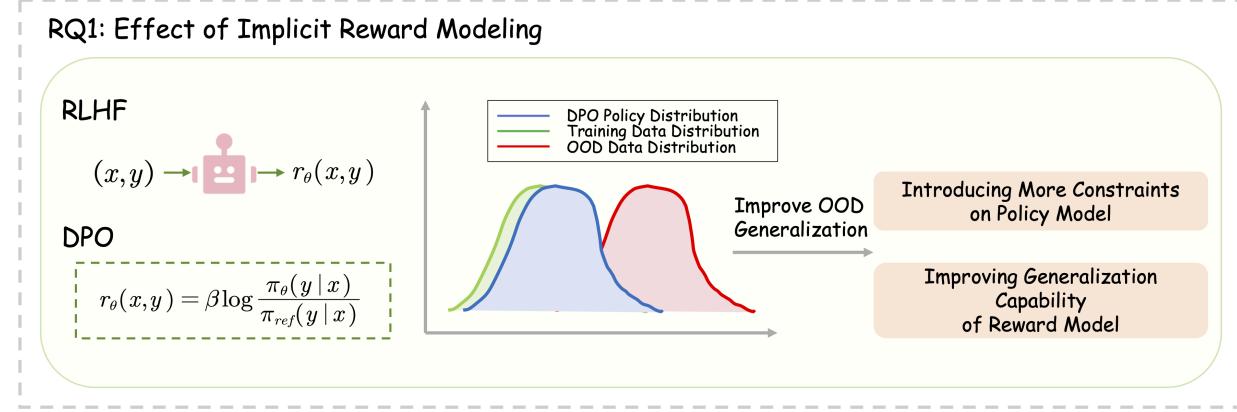


Figure 3: RQ1 Effect of Implicit Reward Modeling. This figure shows that while the DPO policy distribution fits the in-distribution training data well, its implicit reward mechanism leads to poor generalization. Consequently, its performance degrades significantly on out-of-distribution data, revealing a key limitation under distribution shifts compared to explicit reward models.

generalization ability against distribution shifts by regularizing the hidden states. Specifically, they retain the base model’s language model head and incorporate a suite of text-generation losses to preserve the hidden states’ text generation capabilities, while concurrently learning a reward head behind the same hidden states. The two studies primarily focus on explicit reward modeling, whereas strategies to improve OOD generalization in DPO remain unexplored.

Yan et al. [2025] identified the "3D-Properties" in DPO’s implicit reward modeling: Drastic Drop in rejected response likelihood, Degradation into response suppression, and Dispersion effect on unseen responses. Theoretical analysis reveals that gradient imbalance arises from the inverse proportionality between gradient magnitudes of chosen π^+ and rejected π^- responses, causing gradient explosion for π^+ (as $\pi^- \rightarrow 0$) and vanishing gradients for π^+ . Experiments show these issues can be mitigated via adaptive scaling of β or integration with SFT loss, though DPO’s sensitivity to preference data distribution remains a fundamental limitation.

Cen et al. [2025] proposed Value-Incentivized Preference Optimization (VPO), whose technical core is the addition of a regularization term to the DPO loss. This term measures the log-probability gain of the current policy over the original reference policy, and a simple sign switch is used to maximize or minimize its expected value. For online scenarios, this mechanism drives exploration by penalizing overconfidence in known behaviors; for offline scenarios, it suppresses overfitting by rewarding consistency with a trusted data distribution.

We believe that to address the issue of out-of-distribution (OOD) generalization, there are two potential methods: introducing more constraints on policy model optimization or improving the generalization capability of the reward model.

RQ2: Effect of Different Feedback. RLHF leverages point-wise reward from the reward model to optimize the policy model, as well DPO uses point-wise reward and pair-wise preference data since it’s derived from RLHF. However, some studies employ other forms of feedback (e.g. List-wise, Binary, Step-wise, Token-wise, etc) as the reward signal for optimization.

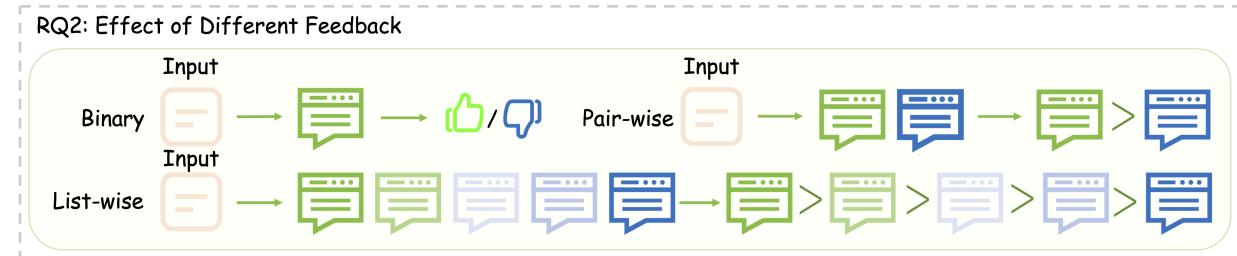


Figure 4: RQ2 Effect of Different Feedback. The figure contrasts three parallel feedback granularities: Binary (absolute judgment), Pair-wise (relative comparison), and List-wise (multi-item ranking).

Dong et al. [2023] proposed a reward ranked fine-tuning method to explore the list-wise feedback. The core idea of RAFT is that the model iteratively learns from the induced best-of-K policy Nakano et al. [2022], Cobbe et al. [2021], which samples K responses and selects the one with the highest reward as the final output. Then the model is fine-tuned on the optimal responses.

Rank Responses to align Human Feedback (RRHF) Yuan et al. [2023] fully exploited rank from human annotators or reward models by combining a modified rank loss with SFT loss. To avoid explicit reward model, they take length-normalized conditional log probability of responses under policy model π_θ as reward score. The core idea is to let the policy model π_θ give larger probabilities for better responses and give smaller probabilities for worse responses.

Yuan et al. [2024c] proposed a data collection method named ULTRAINTERACT for tree-structured preference data, especially in the reasoning domain. Specifically, they decomposed complex tasks into multiple steps to obtain multi-turn model actions. These paired models correct and incorrect actions organized in binary tree structures. They also introduced a critique model to refine the solution while the actor interact with the Python environment. Besides, by training an explicit reward model, they enhanced the Bradley-Terry (BT) objective with a term to directly boost the rewards of chosen actions while decreasing the rewards of rejected ones.

Besides, Ethayarajh et al. [2024] proposed KTO to maximize the utility of LLM generations directly rather than maximizing the log-likelihood of preferences inspired from prospect theory Tversky and Kahneman [1992]. This approach eliminates the need for two preferences for the same input, as it focuses on discerning whether a preference is desirable or undesirable.

Standard RLHF deploys reinforcement learning in a specific token-level MDP, while DPO is derived as a bandit problem in which the whole response of the model is treated as a single arm. Rafailov et al. [2024a] theoretically show that we can derive DPO in the token-level MDP as a general inverse Q-learning algorithm, which satisfies the Bellman equation.

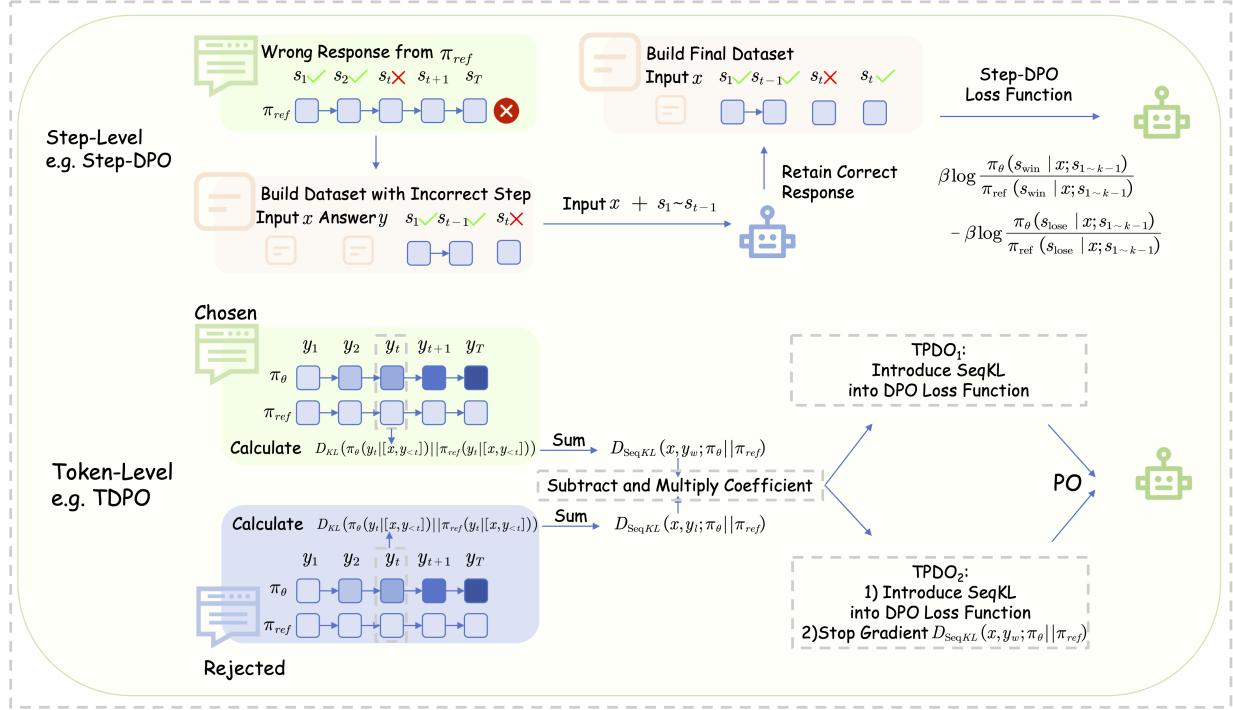


Figure 5: Step-Level and Token-Level Feedback Methods. The upper part shows Step-DPO, which provides step-level supervision by focusing the optimization unit on the first erroneous reasoning step. The lower part depicts TDPO, which performs finer-grained policy optimization by applying a per-token KL divergence constraint.

Lai et al. [2024] proposed Step-DPO to address the limited effectiveness of DPO on long-chain reasoning tasks. They argued that DPO's holistic comparison of entire answers lacks the fine-grained supervision needed to identify specific errors within a long reasoning chain. Step-DPO shifts the fundamental unit of optimization from the entire response to an individual reasoning step. The method precisely targets the first erroneous step in a flawed solution and conditioned on the preceding correct steps, trains the model to prefer a correct, self-generated continuation over the original incorrect

one. This approach provides more precise process-level supervision, enabling the model to accurately locate and correct errors.

Zeng et al. [2024] introduced Token-level Direct Preference Optimization (TDPO), a novel approach to align LLMs with human preferences by optimizing policy at the token level. TDPO incorporates forward KL divergence constraints for each token, improving alignment and diversity. Utilizing the Bradley-Terry model for a token-based reward system, TDPO enhances the regulation of KL divergence, while preserving simplicity without the need for explicit reward modeling.

Previous research derived pairwise preferences using pointwise rewards and the BT model. However, this approach was not comparable to direct pairwise preference modeling and failed to address inconsistencies within pairwise preferences. To overcome these limitations, some recent studies Wu et al. [2024b], Rosset et al. [2024], Munos et al. [2024] have introduced Nash learning methodologies where each player is an LLM that outputs responses and aims to maximize its probability of being preferred over its opponent.

Inspired by generative adversarial networks (GAN) Goodfellow et al. [2014], Self-Play fine-tuning (SPIN) considered a two-player game Wu et al. [2024b], Chen et al. [2024a], where the main player distinguishes the generated responses are from model or human, while the opponent player generates responses indistinguishable from human. This approach also eliminated the need for a reward model and derived a objective in a similar form to DPO.

An intuitive interpretation of DPO would lead one to believe it increases the likelihood of chosen responses while decreasing the likelihood of rejected responses. However, this does not support a well-observed phenomenon in which the likelihood of the chosen responses actually decreases over time Pal et al. [2024], Rafailov et al. [2024b]. To address such issues, some works focus Pang et al. [2024], Yuan et al. [2023], Xu et al. [2024b] on introducing extra objective terms.

Yuan et al. [2023] added a negative log-likelihood (NLL) loss similar to SFT (supervised fine-tuning) to rank loss. By doing so, They force the model to learn the response with the highest reward. Xu et al. [2024b] incorporated a behavior cloning (BC) regularizer Hejna et al. [2024] to ensure that π_θ does not deviate from the preferred data distribution. They prove that the regularizer can boil down to adding a NLL term on the preferred data distribution.

Pang et al. [2024] have found that when training with DPO without negative log-likelihood (NLL) loss, the log probabilities of chosen sequences barely increase over training; when training with DPO with NLL loss normalized by the total response length, the log probabilities increase noticeably. Thus, they believed that NLL enhances learning over the winning response from each pair.

Lin et al. [2024d] proposed the cDPO algorithm, which improved reasoning capabilities by identifying and penalizing critical tokens. cDPO employed contrastive estimation to identify these negative-exclusive critical tokens efficiently. Departing from the example-level implicit reward mechanism of DPO, cDPO innovatively reframed it into a token-level explicit reward function, applying dynamic penalty weights to each token in negative trajectories. This systematically reduced the probability of generating highly critical tokens. Experiments demonstrated that cDPO significantly improved model accuracy on mathematical reasoning tasks.

Xiong et al. [2025] proposed Multi-turn Direct Preference Optimization (M-DPO) to address multi-turn mathematical reasoning tasks that integrate external tools. The approach formulates the problem as an MDP to theoretically re-parameterize the trajectory-level reward as a sum of log-probabilities over the agent's actions. M-DPO implements this by "masking out" external observations during loss computation, forcing the model to focus solely on optimizing its own reasoning and execution steps. This prevents the model from learning to predict uncontrollable environmental messages.

RQ3: Effect of KL Penalty Coefficient and Reference Model. As discussed in RQ1, DPO implicitly learns a reward model r_θ given an input x and an output $r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$, where π_θ and π_{ref} denote the distributions parameterized by the fine-tuned LLM and the reference LLM, respectively, and β controls the strength of the Kullback-Leibler (KL) divergence regularization applied from the reference LLM. Some recent studies Wu et al. [2024a], Liu et al. [2024c], Rafailov et al. [2024a], Xu et al. [2024b], Hong et al. [2024], Meng et al. [2024] have investigated the impact of the KL penalty coefficient β and the choice of reference model π_{ref} .

In KL-constrained RL and DPO, the KL penalty coefficient controls the trade-off between maximizing the reward and minimizing the deviation from the reference policy. Wu et al. [2024a] introduces a novel framework that dynamically calibrates β at the batch level, informed by data quality considerations. Additionally, this method incorporates β -guided data filtering to safeguard against the influence of outliers. Liu et al. [2024c] have explored the optimal KL constraint strength for DPO, finding that a smaller KL constraint generally improves performance until the constraint becomes too small and leads to performance degradation. Empirically, following token log-probability difference experimental settings Rafailov et al. [2024a], results reveal that as the strength of the KL constraint decreases, the DPO-fine-tuned

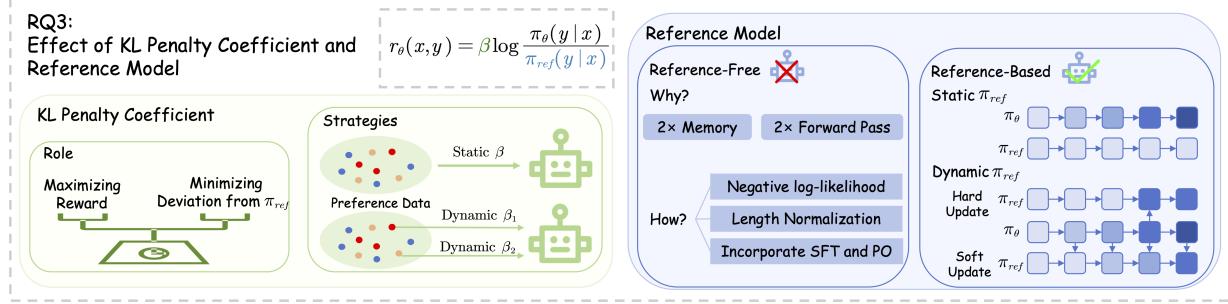


Figure 6: RQ3 Effect of KL Penalty Coefficient and Reference Model. The left panel illustrates the key role of the KL penalty in controlling the trade-off between reward maximization and policy deviation, showcasing both static (fixed coefficient) and dynamic (adaptive coefficient) strategies. The right panel categorizes methods based on their reliance on a reference model into two main classes: reference-based and reference-free approaches.

model begins to assign significantly different probabilities to a small subset of specific tokens compared to the reference model.

Given that the reference model requires double the memory and forward passes, recent research Xu et al. [2024b], Meng et al. [2024] has explored alternative forms of regularization to replace the reference model. Besides, we usually initialize the reference model with a supervised fine-tuning (SFT) model, which involves a two-stage training process(i.e. SFT and preference optimization). Empirical evidence suggests that the effectiveness of DPO has strong reliance on the training effect of the LLMs after SFT Feng et al. [2024]. To address this issue, some studies integrate SFT into preference optimization Hong et al. [2024].

Xu et al. [2024b] argues that DPO Rafailov et al. [2023] is memory-inefficient and speed-inefficient due to twice memory for reference and policy model and twice forward pass. To solve this issue, their proposed method Contrastive Preference Optimization (CPO) removes the reference model term by proving the upper boundary of DPO loss. Furthermore, CPO incorporates a behavior cloning (BC) regularizer (i.e. negative log-likelihood NLL) to ensure that π_θ does not deviate from the preferred data distribution.

Hong et al. [2024] observed that in SFT stage, the absence of penalty in Cross-Entropy loss hinders human preference learning. Inspired by unlikelihood penalty Welleck et al. [2019], they incorporate an odds ratio-based penalty to negative log-likelihood (NLL), where $\text{odds}_\theta(y|x) = \frac{\pi_\theta(y|x)}{1-\pi_\theta(y|x)}$. Consequently, they proposed ORPO to incorporate SFT and preference optimization by the guidance of NLL and odds ratio (OR), where $\text{OR}_\theta(y_w, y_l) = \frac{\text{odds}_\theta(y_w|x)}{\text{odds}_\theta(y_l|x)}$, which requires neither an SFT warm-up stage nor a reference model. Besides, Compared to the probability ratio (PR), $\text{PR}_\theta(y_w, y_l) = \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}$, OR empirically avoids an overly extreme optimization to tokens in rejected response.

However, Liu et al. [2024c] have investigated the necessity of the reference model, and reveal that the KL constraint from the reference model in DPO helps to stabilize the model behavior. Furthermore, a stronger reference model in DPO finetuning can improve DPO’s effectiveness with a larger KL penalty coefficient. Meng et al. [2024] have proposed SimPO where the reference model is eliminated without theoretical analysis. They also have found that the logarithmic ratio between the policy model and the reference model can serve to implicitly counteract length bias.

Additionally, Xu et al. [2024a] found that DPO might discover solutions exploiting out-of-distribution data, posing a risk of deviating excessively from the reference model even when the reference model aligns well with human preferences, which reveals the potential risk of the reference model regularisation.

Gorbatovski et al. [2025] proposed a Trust Region DPO (TR-DPO) to address overoptimization in offline alignment by dynamically updating the reference policy during training. The method incorporates soft updates (interpolating the current policy with the reference policy) and hard updates (periodically replacing the reference policy) to reduce reliance on out-of-domain data, enabling higher policy deviation while maintaining generation quality. Experiments demonstrated that TR-DPO significantly outperforms standard DPO in dialogue and summarization tasks.

Lee et al. [2025] proposed ϵ -DPO, which dynamically adjusted instance-level KL penalties through perturbation strategies. By observing the monotonicity of log-likelihood ratios between chosen and rejected responses under perturbed β values, the algorithm adaptively determined KL coefficients without relying on batch-level statistics or reference policy updates.

Xie et al. [2025] proposed Exploratory Preference Optimization (XPO), an enhanced variant designed for online DPO settings. This work provides a profound insight that DPO implicitly performs a form of Q^* -function approximation. Building on this observation, XPO ingeniously augments the standard DPO objective with an exploration bonus derived from the "optimism in the face of uncertainty" principle in reinforcement learning, encouraging the model to deliberately explore novel behaviors beyond the support of the initial reference policy. Through this simple modification, XPO theoretically overcomes the sample inefficiency of conventional online DPO caused by "passive exploration."

RQ4: Online DPO. RLHF Ouyang et al. [2022a] is an online framework for alignment, while recent advances in offline methods (e.g. DPO Rafailov et al. [2023], Xu et al. [2024e]) show empirically efficient in practice. Online algorithms tend to be more computationally intensive than offline algorithms, due to sampling and training a reward model. Thus some studies have provided insights into the performance gap between online and offline algorithms.

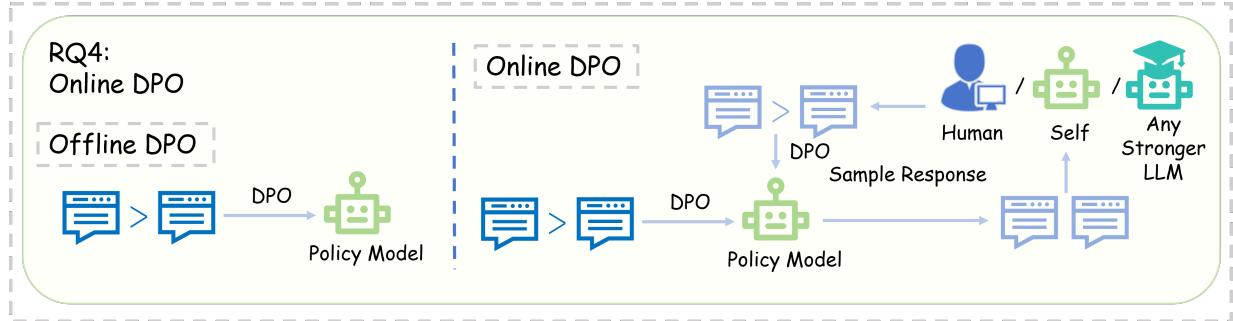


Figure 7: RQ4 Online DPO. This figure shows the training processes for Offline DPO and Online DPO. In the Online DPO, newly sample responses from each iteration can be annotated through various means, including by human annotators, the model itself, or any other stronger model.

Tang et al. [2024] have found empirically that online algorithms seem to generally achieve a better trade-off compared to offline algorithms. Concretely, with the same budget on the KL divergence, online algorithms obtain generally better performance than offline algorithms.

Wang et al. [2024a] claimed that current offline RLHF only reflect the "ordinal relationship" of candidate responses, neglecting the extent to which the optimal response is preferred over others. To address this issue, they proposed a metric called *reward difference coefficients* to reweight preferences. Additionally, they developed a *difference model* to predict these coefficients, providing more accurate supervision signals for offline methods.

To explore the online algorithm of DPO, iterative and online DPO have been implemented, raising the intriguing question of how to efficiently collect new preference datasets. Besides, since the model evolves over training, DPO is inevitably off-policy. Some recent research has delved into iterative DPO.

Yuan et al. [2024d] introduced the concept of Self-Rewarding to simultaneously enhance generation and reward performance, which consists of two stages: "Instruction Following" and "Self Rewarding". During the "Instruction Following" training, DPO was used to train the LLM to align with the instruction following dataset. In the "Self Rewarding" phase, candidate responses were evaluated by the model itself (LLM-as-a-judge Zheng et al. [2023a]), and each candidate was assigned a score considering five metrics (i.e. relevance, coverage, usefulness, clarity, and expertise) to construct a preference dataset for preference optimization. For the next iteration, the fine-tuned model acts as the reference model. The two stages are iteratively performed until performance degradation.

Xu et al. [2024c] proposed Pairwise Cringe Optimization (PCO) to generalize Cringe Loss Adolphs et al. [2022] which is applied to binary feedback to pairwise feedback setting by using a soft margin extension. Besides, they also apply PCO iteratively in a way similar to Yuan et al. [2024d] outperforming DPO training in the same iterative way, while iterative PCO requires the reward mode to label generated responses.

Guo et al. [2024a] observed that LLMs can approximate well human labelling and can generate reliable preferences over responses, thus potentially could alleviate the distribution shift issue on reward models. Concurrently to Self-Rewarding Yuan et al. [2024d], Guo et al. [2024a] proposed an iterative DPO method called online AI feedback (OAIF) similar to Self-Rewarding. While Self-Rewarding obtains feedback from itself, OAIF can leverage feedback from any LLM, including stronger ones.

Additionally, Kim et al. [2024] proposed a simple yet effective step-wise method (sDPO) for better alignment. They split the preference dataset into partitions and then employed DPO on the partitions iteratively while placing the policy

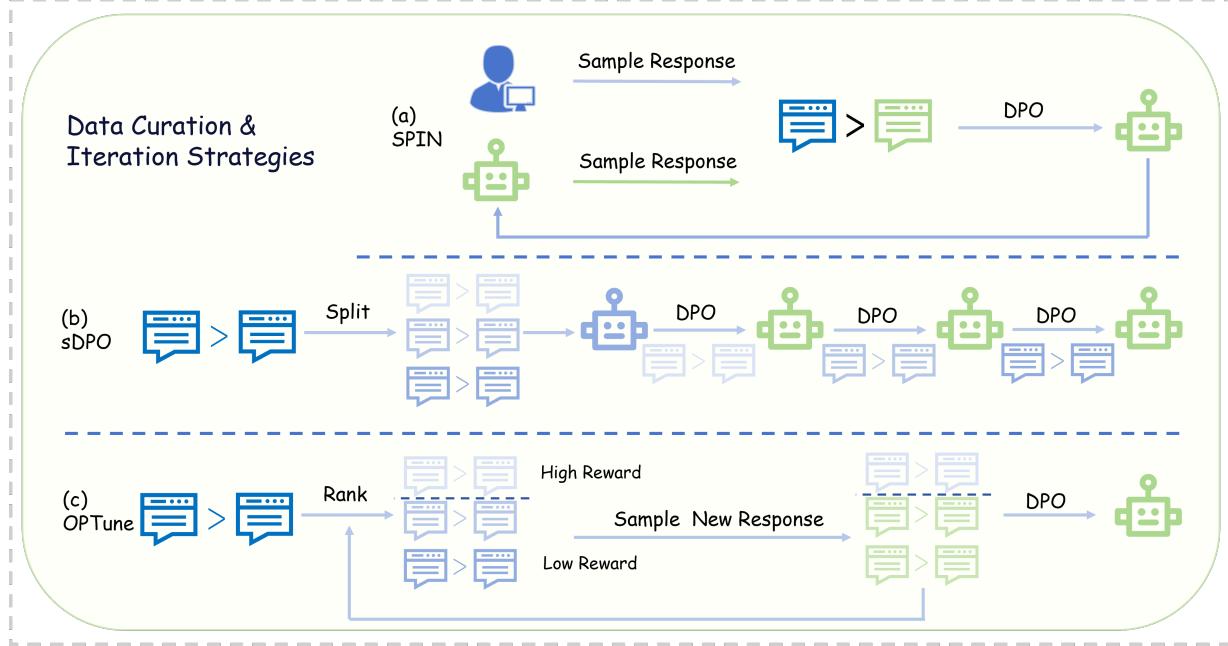


Figure 8: Data Curation and Iteration Strategies in Online DPO. The panel (a) shows a self-play framework that iteratively optimizes by pairing model responses against human responses. The panel (b) shows sDPO, which trains on data partitions with a dynamically updated reference model. The panel (c) shows the OPTune algorithm, which combines selective data regeneration with a weighted-DPO loss to accelerate optimization.

model of the current iteration as the reference model of the next iteration. They believed this step-wise DPO set a tighter low bound for policy model optimization.

Pang et al. [2024] proposed iterative reasoning DPO, especially for the COT reasoning task. They trained a variant of DPO that includes a negative log-likelihood (NLL) loss term for the chosen response, which proves crucial for performance. Given the newly trained model, they then iterate the procedure by generating new pairs, and training again, starting from the previously trained model. They find that reasoning performance improves over multiple iterations until it eventually saturates.

Chen et al. [2024b] introduced an efficient data generation algorithm (OPTune) for online RLHF. Since they observed that in RLHF generating responses accounts for approximately 70% time-consume and resampling low reward responses should be more beneficial for consequent training, they selectively regenerate the lowest-rewarded responses. Furthermore, they utilized a variant of iterative DPO objective named w-DPO that assigns greater weight to preference pairs with larger reward gaps to accelerate preference optimization.

Inspired by a phenomenon in nature that intraspecific competition drives species evolution, Qi et al. [2024] proposed an Online Fast-Slow chasing DPO (OFS-DPO) for preference alignment. Specifically, they introduced two identical LoRA(Low-rank Adaptive) Hu et al. [2021] modules with different optimization speeds using DPO with a new regularization term. Besides, in every certain step, they swap the roles of the fast module and slow module to maintain a more sustained gradient update momentum. This method has been validated by their theoretical analysis and empirical results.

Chen et al. [2024d] proposed an iterative self-play framework to fine-tune the model in a supervised way. In each round, new preference pairs are constructed by taking the responses sampled from the model as losers and the responses from humans as winners. Then DPO is applied in each round iteratively to optimize the model.

Bose et al. [2025] proposed Hybrid Preference Optimization (HPO) to address the limitations of traditional DPO in sample efficiency and exploration capability by integrating offline preference data with online exploration mechanisms. The method theoretically established sample complexity bounds for policy optimization by relaxing concentrability conditions on offline data and introducing an optimistic regularizer for online feedback, which proved faster convergence rates than offline DPO or online exploration methods. HPO combined offline data and online exploration through an objective function that included a DPO loss on offline data and an exploration-aware regularizer, where parameter γ balances offline-online contributions to approach optimal policies under limited samples.

RQ5: Reward Hacking. A critical issue of both RLHF and DPO is reward hacking. This occurs when the policy achieves a high reward but fails to meet the actual objectives. The root cause of this issue is that the explicit or implicit Reward Model (RM) is not a perfect proxy for human preferences and exhibits limited generalization to out-of-distribution (OOD) scenarios. In contrast, the policy can learn to generate OOD examples to exploit these weaknesses. Consequently, the policy model might develop specific response patterns to exploit the RM.

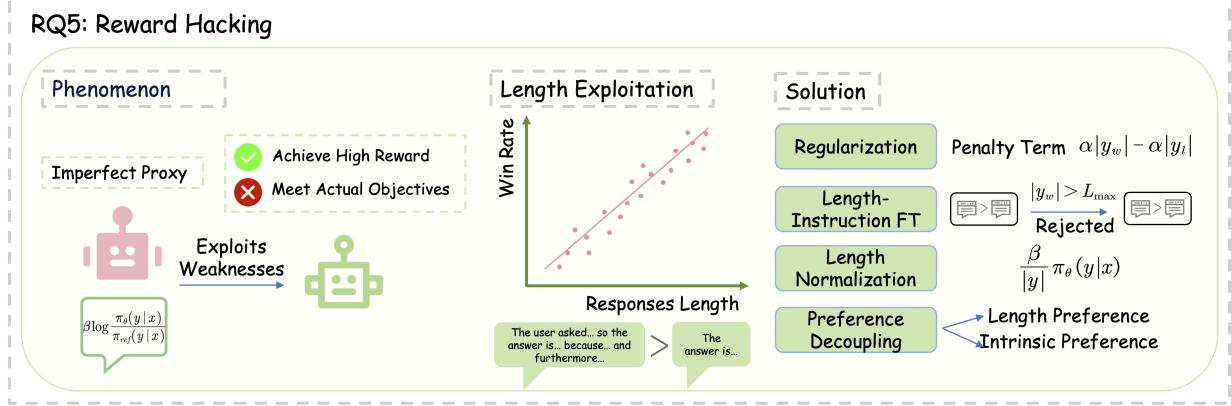


Figure 9: RQ5 Reward Hacking. Reward hacking occurs when a policy exploits the limited generalization of its reward model to achieve a high reward without meeting the actual objectives. A prominent example is length exploitation, where models generate longer responses to exploit data biases, often at the cost of quality. This figure shows various mitigation strategies to counter this issue, such as length normalization and regularization.

For instance, some recent research Dubois et al. [2024], Singhal et al. [2023], Kabir et al. [2024] find that both humans and models tend to have a *length exploitation* whereby they prefer longer responses over shorter ones in pairwise preferences. For example, prior work Wang et al. [2023b] found that, given head-to-head model comparisons, there is a correlation of 0.96 between win rates and the average number of unique tokens in the model’s responses. Moreover, with such bias in preferences, aligned models tend to generate longer responses with potential hallucinations or repetitions.

Park et al. [2024] investigated disentangling verbosity from quality in DPO. They investigated length exploitation in the DPO setting and linked it to out-of-distribution bootstrapping. Besides, they employed a principled but simple regularization derived from RLHF named Regularization-DPO (R-DPO), which prevents length exploitation while still maintaining improvements in model quality.

Yuan et al. [2024b] introduced a novel method named Length-Instruction Fine-Tuning (LIFT-DPO). This method incorporates length constraint instructions into general instruction datasets and augments original preference pairs by rejecting chosen responses that exceed the specified length limit. Empirical results indicate that LIFT-DPO shows no performance degradation on standard benchmarks and performs well on length instruction benchmarks.

SimPO Meng et al. [2024] modified the DPO loss to align the training objective with the generation objective, which also introduced a length normalization. Their empirical results demonstrated that length normalization increases the reward difference across all preference pairs, irrespective of their length. Additionally, they found that removing length normalization results in a strong positive correlation between reward and response length, leading to length exploitation.

Liu et al. [2024d] investigated the DPO optimization objective, uncovering a significant link between its implicit reward and the length of data. This connection inadvertently steers the optimization process, resulting in length-sensitive training and verbose outputs. To address this issue, the researchers introduced LD-DPO, a novel length-desensitization enhancement for DPO, which seeks to mitigate DPO’s sensitivity to data length by separating the relatively minor explicit length preference from other implicit preferences. By doing so, LD-DPO focuses on capturing intrinsic preferences instead of length preferences. The theoretical analysis conducted by Liu et al. [2024d] forms the foundation for this improvement, addressing the verbosity problem that arises during DPO training.

Rashidinejad and Tian [2025] identified that conventional DPO suffers from two types of reward hacking: Type I arises from over-optimizing out-of-distribution actions due to partial data coverage, while Type II stems from performance degradation of the initial model when preference data poorly covers high-reward actions. They proposed POWER-DL, which combined robust reward optimization (POWER) with dynamic label weighting. POWER introduced weighted entropy regularization to enforce pessimism on uncovered actions, while dynamic labels adaptively suppressed destructive updates to the initial policy. This approach required only objective function modifications without additional model training or data augmentation.

RQ6: Alignment Tax. RLHF aims to align models with human preferences. However, prior studies Ouyang et al. [2022a], Tu et al. [2023], OpenAI [2024a], Anthropic [2024], Wu et al. [2024c] have found a phenomenon known as the *alignment tax*. This phenomenon first refers to the improved preference accompanies degraded performances on other NLP tasks after alignment Ouyang et al. [2022a]. Kim and Seo [2024] observed that the key neurons for safety and helpfulness significantly overlap, but they require different activation patterns of the shared neurons, which could account for the alignment tax. Therefore alignment tax poses a significant challenge to achieving effective and practical model alignment. Consequently, some studies Lu et al. [2024a], Lin et al. [2024a], Fu et al. [2024], Lou et al. [2024a], Pentyala et al. [2024] have investigated the alignment tax and proposed methods to reduce its effect, focusing on *data iteration*, *model merging (averaging)*.

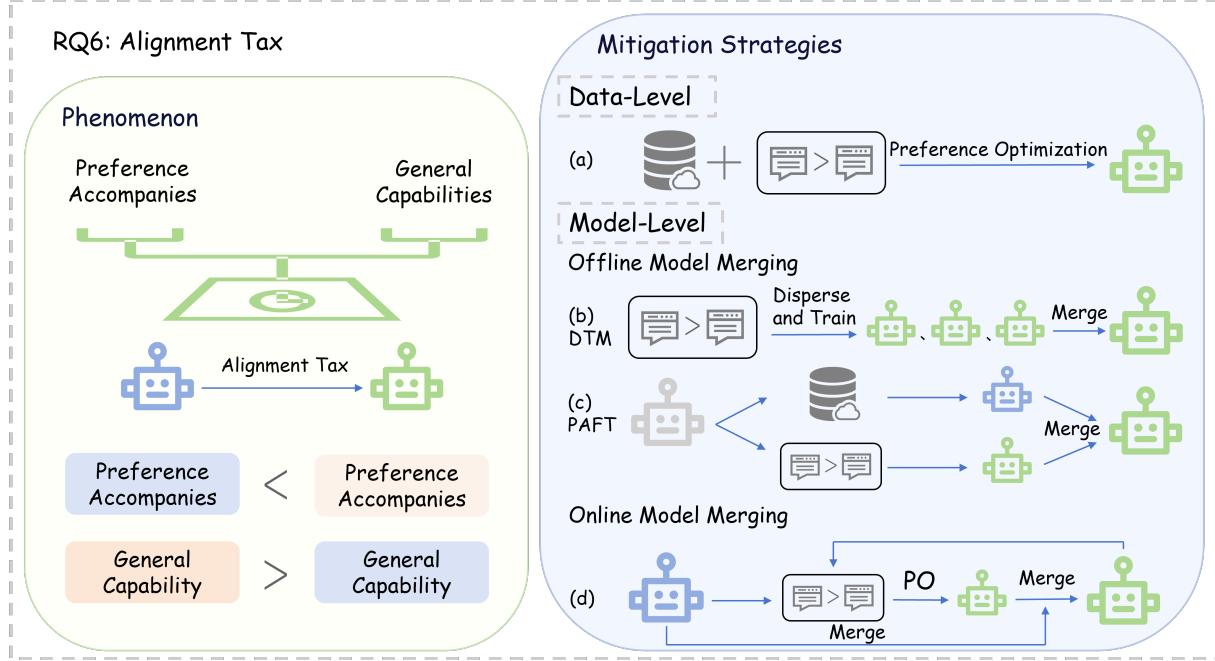


Figure 10: RQ6 Alignment Tax. The scale in the left panel illustrates the "alignment tax"—the trade-off between preference accompanies and general capabilities. The right panel outlines various strategies to mitigate this issue, primarily focusing on data-level and model-level strategies.

Ouyang et al. [2022a] proposed mitigating the alignment tax by incorporating pretraining data into the RLHF fine-tuning process. However, this approach does not entirely eliminate performance regressions and may increase the likelihood of certain undesirable behaviors if these behaviors are present in the pretraining data.

Fu et al. [2024] have found that there is also an alignment tax phenomenon present during the SFT stage. They hypothesize that data biases are likely one of the causes. To address the issue, they proposed to disperse the instruction following data into portions and train multiple sub-models using different data portions. Finally, they merged multiple models into a single one via model merging techniques to reduce alignment tax.

Lin et al. [2024a] conducted experiments with existing RLHF algorithms including DPO, which validated an obvious alignment tax in NLP tasks. They explored model averaging, which interpolates between pre and post RLHF model weights, to achieve a more efficient reward-tax Pareto front.

Compared to offline model merging methods Lin et al. [2024a], Lu et al. [2024a] proposed an online method where at each optimization step of RLHF, they online merged the gradients of the policy model with the delta parameters of the SFT model. Their experimental findings indicate that this online approach is more effective than offline merging methods in reducing alignment tax while maximizing rewards.

Pentyala et al. [2024] introduced PAFT to address the issue, a new parallel training paradigm that independently performs SFT and DPO (or other preference alignment) with the same pre-trained model on respective datasets. The model produced by SFT and the model from preference alignment are then merged into a final model by parameter fusing.

Lou et al. [2024a] have found that alignment tax will accumulate in the iterative fine-tuning of each dimension, resulting in misalignment on previous dimensions or even catastrophic model collapse. To address the issue, they proposed Sequential Preference Optimization (SPO) which introduces additional constraints to the optimization problem to prevent model performance degradation.

Related to alignment tax, *multi objective learning* involves trade-off where improvements in alignment with one objective (e.g., helpfulness) can lead to a decrease in performance in other objectives (e.g., safety). Guo et al. [2024b] introduced Controllable Preference Optimization (CPO), a method that assigns explicit preference scores to different objectives, thereby guiding the model to generate responses that meet specified requirements. Their empirical results shows that this approach mitigates the impact of the alignment tax and achieves Pareto improvements in multi-objective alignment.

Zhou et al. [2024a] proposed Multi-Objective Direct Preference Optimization (MODPO), an extension of DPO for multiple alignment objectives. Specifically, compared to DPO loss, MODPO loss includes a margin to steer language models by multiple objectives. Empirical results in safety alignment and long-form QA show that MODPO produced a Pareto front catering to diverse preferences.

Xiao et al. [2025] identified that DPO alignment introduced significant calibration errors in large language models, causing overconfidence in predictions and degrading reliability on tasks like TruthfulQA. They proposed RCFT, which integrated dynamic confidence adjustment mechanisms during calibration-aware fine-tuning and employed temperature scaling during inference to align model output probabilities with ground truth. This approach required only loss function modifications without additional data or constraints. Experiments demonstrated RCFT reduces calibration errors on models while maintaining DPO's alignment performance.

Method	Objective
RQ0: why DPO?	
PPO Ouyang et al. [2022a]	$r_\phi(x, y) - \beta \mathbb{D}_{\text{KL}}[\pi_\theta(y x) \parallel \pi_{\text{ref}}(y x)]$
DPO Rafailov et al. [2023]	$-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} - \beta \log \frac{\pi_\theta(y_t x)}{\pi_{\text{ref}}(y_t x)} \right)$
RQ1: Generalization Ability of RM?	
IPO Azar et al. [2023]	$\left(\beta \log \frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} - \beta \log \frac{\pi_\theta(y_t x)}{\pi_{\text{ref}}(y_t x)} - \frac{1}{2} \right)^2$
VPO Cen et al. [2025]	$\begin{aligned} & \left\{ -\sum_{s=1}^t \log \sigma \left(\beta \log \frac{\pi(y_+^{(s)} x^{(s)})}{\pi_{\text{ref}}(y_+^{(s)} x^{(s)})} - \beta \log \frac{\pi(y_-^{(s)} x^{(s)})}{\pi_{\text{ref}}(y_-^{(s)} x^{(s)})} \right) \right. \\ & \left. + \alpha \beta \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{cal}}(\cdot x)} [\log \pi(y x) - \log \pi_{\text{ref}}(y x)] \right\} \end{aligned}$
RQ2: Rewarding Signals?	
RRHF Yuan et al. [2023]	$\begin{aligned} & \sum_{r_i < r_j} \max(0, p_i - p_j) - \sum_t \log P_\pi(y_{i,t} x, y_{i,<t}) \\ & \text{where } p_i = \frac{\sum_t \log P_\pi(y_{i,t} x, y_{i,<t})}{\ y_i\ }, i' = \arg \max_i r_i \end{aligned}$
SPIN Chen et al. [2024a]	$\begin{aligned} & \ell \left(\lambda \log \frac{p_\theta(y x)}{p_{\theta_t}(y x)} - \lambda \log \frac{p_\theta(y' x)}{p_{\theta_t}(y' x)} \right) \\ & \text{where } \ell(t) := \log(1 + \exp(-t)), y \sim p_{\text{data}}(\cdot x), y' \sim p_{\theta_t}(\cdot x) \end{aligned}$
Step-DPO Lai et al. [2024]	$-\log \sigma(\beta \log \frac{\pi_\theta(s_w x, s_{1 \sim k-1})}{\pi_{\text{ref}}(s_w x, s_{1 \sim k-1})} - \beta \log \frac{\pi_\theta(s_t x, s_{1 \sim k-1})}{\pi_{\text{ref}}(s_t x, s_{1 \sim k-1})})$
T-DPO Zeng et al. [2024]	$\begin{aligned} & -\log \sigma \left(\beta \log \frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} - \beta \log \frac{\pi_\theta(y_t x)}{\pi_{\text{ref}}(y_t x)} \right) \\ & -\beta (D_{\text{SeqKL}}(x, y; \pi_{\text{ref}} \pi_\theta) - D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \pi_\theta)) \end{aligned}$
KTO Ethayarajh et al. [2024]	$\begin{aligned} & -\lambda_w \sigma \left(\beta \log \frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} - z_{\text{ref}} \right) - \lambda_t \sigma \left(z_{\text{ref}} - \beta \log \frac{\pi_\theta(y_t x)}{\pi_{\text{ref}}(y_t x)} \right), \\ & \text{where } z_{\text{ref}} = \mathbb{E}_{(x, y) \sim \mathcal{D}} [\beta \text{KL}(\pi_\theta(y x) \pi_{\text{ref}}(y x))] \end{aligned}$
cDPO Lin et al. [2024d]	$\begin{aligned} & \ell_{\text{cDPO}} = -\sum_{i=1}^M \log \sigma(\phi(x_i, y_i^n) - \phi_s(x_i, y_i^n, s_i^n)) \\ & \text{where } \phi_s(x_i, y_i^n, s_i^n) = \gamma \sum_{t=1}^T (1 - s_t^n) \cdot \log \frac{\pi_\theta(y_t^n x_i, y_{<t}^n)}{\pi_{\text{ref}}(y_t^n x_i, y_{<t}^n)} \\ & \text{and } s_t^n = \frac{P_p(y_t^n x_i, y_{<t}^n)^{1+\beta}}{P_n(y_t^n x_i, y_{<t}^n)^{\beta} \cdot Z} \end{aligned}$
M-DPO Xiong et al. [2025]	$\log \sigma \left(\eta \sum_{h=1}^H \left[\log \frac{\pi_{\theta,h}(a_h s_h^i)}{\pi_{\text{ref},h}(a_h s_h^i)} - \log \frac{\pi_{\theta,h}(a_h^w s_h^w)}{\pi_{\text{ref},h}(a_h^w s_h^w)} \right] \right)$
RQ3: β Coefficient and Reference Model?	
β -DPO Wu et al. [2024a]	$\begin{aligned} & -\log \sigma \left(\beta_i \log \frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} - \beta_i \log \frac{\pi_\theta(y_t x)}{\pi_{\text{ref}}(y_t x)} \right) \\ & \text{where } \beta_i = [1 + \alpha(M_i - M_0)]\beta_0, \\ & \text{and } M = \beta_0 \log \left(\frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} \right) - \beta_0 \log \left(\frac{\pi_\theta(y_t x)}{\pi_{\text{ref}}(y_t x)} \right) \end{aligned}$
CPO Xu et al. [2024b]	$-\log p_\theta(y_w x) - \log \sigma(\beta \log \pi_\theta(y_w x) - \beta \log \pi_\theta(y_t x))$
ORPO Hong et al. [2024]	$\begin{aligned} & -\log p_\theta(y_w x) - \lambda \log \sigma \left(\log \frac{p_\theta(y_w x)}{1 - p_\theta(y_w x)} - \log \frac{p_\theta(y_t x)}{1 - p_\theta(y_t x)} \right), \\ & \text{where } p_\theta(y x) = \exp \left(\frac{1}{ y } \log \pi_\theta(y x) \right) \end{aligned}$
XPO Xie et al. [2025]	$\alpha \sum_{i=1}^t \log \pi(\tilde{\tau}^{(i)}) - \sum_{(\tau_+, \tau_-) \in \mathcal{D}_{\text{pref}}^{(i)}} \log \left[\sigma \left(\beta \log \frac{\pi(\tau_+)}{\pi_{\text{ref}}(\tau_+)} - \beta \log \frac{\pi(\tau_-)}{\pi_{\text{ref}}(\tau_-)} \right) \right]$
RQ4: Training Strategy of DPO?	
OPTune Chen et al. [2024b]	$\begin{aligned} & -R(x, y_w, y_t) \cdot \log \sigma \left(\beta_1 \log \frac{\pi_{t+1}(y_w x)}{\pi_t(y_w x)} - \beta_1 \log \frac{\pi_{t+1}(y_t x)}{\pi_t(y_t x)} \right) \\ & \text{where } R(x, y_w, y_t) = \sigma[\beta_2(r(x, y_w) - r(x, y_t))] \end{aligned}$
IRPO Pang et al. [2024]	$-\log \sigma \left(\beta \log \frac{\pi_\theta(c_w, y_w x_i)}{\pi_\theta(c_w, y_w x_i)} - \beta \log \frac{\pi_\theta(c_t, y_t x_i)}{\pi_\theta(c_t, y_t x_i)} \right) - \alpha \frac{\log \pi_\theta(c_w, y_w x_i)}{ c_w + y_w }$
RQ5: Reward Hacking?	
R-DPO Park et al. [2024]	$-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} - \beta \log \frac{\pi_\theta(y_t x)}{\pi_{\text{ref}}(y_t x)} - (\alpha y_w - \alpha y_t) \right)$
LD-DPO Liu et al. [2024d]	$\begin{aligned} & -\log \sigma \left(\beta_i \log \frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} - \beta_i \log \frac{\pi_\theta(y_t x)}{\pi_{\text{ref}}(y_t x)} \right) \\ & \text{where } \pi_\theta(y x) = \prod_{i=1}^t p^\alpha(y_i x, y_{<i}) \prod_{i=1}^t p^{1-\alpha}(y_i x, y_{<i}) \end{aligned}$
SimPO Meng et al. [2024]	$-\log \sigma \left(\frac{\beta}{ y_w } \log \pi_\theta(y_w x) - \frac{\beta}{ y_t } \log \pi_\theta(y_t x) - \gamma \right)$
POWER-DL Rashidinejad and Tian [2025]	$\begin{aligned} & \mathbb{E}_{\mathcal{D}} [l_i^t L_{\text{POWER}}(x, y_i^+, y_i^-) + (1 - l_i^t) L_{\text{POWER}}(x, y_i^-, y_i^+)] \\ & \text{where } L_{\text{POWER}}(x, y^+, y^-) := \log \left(\sigma \left(\beta \left[w(y^+) \log \pi_\theta(y^+ x) - w(y^-) \log \pi_\theta(y^- x) + w(y^+) - w(y^-) \right] \right) \right. \\ & \left. + \eta \beta w(y^+) \log \pi(y^+ x) \right) \end{aligned}$
RQ6: Alignment Tax?	
SPO Lou et al. [2024a]	$\begin{aligned} & -\log \sigma(\xi_2 \phi_2(x, y_w, y_t) - \xi_1 \phi_1(x, y_w, y_t)) \\ & \text{where } \forall i \in \{1, 2\}, \phi_i(x, y_w, y_t) = \log \frac{\pi_i(y_t x)}{\pi_{i-1}(y_w x)} - \log \frac{\pi_i(y_w x)}{\pi_{i-1}(y_t x)} \end{aligned}$

Table 1: Part variants of DPO objectives discussed in RQs. Note that a method may be discussed across multiple RQs. The table is inspired from Meng et al. [2024].

4 Datasets

Datasets	Task Description	Modality	Dataset Size (#Rows)
Human Labeled			
Summarize from Human Feedback Stiennon et al. [2022]	Summarization	Text	193,841
OpenAI:Webgpt Nakano et al. [2022]	Question Answering	Text	19,578
Nvidia/HelpSteer Wang et al. [2023c]	Preference Tuning	Text	37,120
Nvidia/HelpSteer2 Wang et al. [2024b]	Preference Tuning	Text	21,362
Nvidia/HelpSteer3 Wang et al. [2025a]	Preference Tuning	Text	40,476
OpenAssistant/oasst1 Köpf et al. [2023]	Conversation	Text	88,838
Stanfordnlp/SHP Ethayarajh et al. [2022]	Preference Tuning	Text	385,563
Tasksource/Dpo-pairs Sileo [2024]	Preference Tuning	Text	5,128,939
RLHF-V-Dataset Yu et al. [2024a]	Preference Tuning	Multimodal	5,733
HH-RLHF Bai et al. [2022a]	Preference Tuning	Text	169,352
MM-RLHF Zhang et al. [2025]	Preference Tuning	Multimodal	16,342
Chatbot Arena Zheng et al. [2023a]	Model Evaluation	Text	33,000
Alpaca Farm Human Dubois et al. [2024]	Instruction Tuning	Text	101,000
MMSafe-PO Li et al. [2025a]	Safety Preference	Multimodal	5,667
AI Labeled			
RLAIF-V-Dataset Yu et al. [2024b]	Preference Tuning	Multimodal	33,835
Math-Step-DPO Lai et al. [2024]	Mathematical Reasoning	Text	10,795
UltraFeedback Cui et al. [2023a]	Preference Tuning	Text	63,967
Distilabel-capybara-dpo Argilla [2024]	Conversation	Text	7,563
Chatml-dpo-pairs Mukherjee et al. [2023]	Instruction Tuning	Text	12,859
Magpie-Llama-3.1-DPO Xu et al. [2024d]	Instruction Tuning	Text	100,000
Tldr-preference-dpo davanstrien [2024]	Summarization	Text	522
Zsql-postgres-dpo Zerolink [2024a]	Text-to-SQL	Text	259,326
Truthy-dpo Zerolink [2024b]	Truthfulness Preference	Text	1,016
Py-dpo Jondurbin [2024a]	Code Generation	Text	9,466
MMInstruction/VLFeedback Li et al. [2023a]	Instruction Tuning	Multimodal	80,258
sqrti/SPA-VL Zhang et al. [2024a]	Safety Preference	Multimodal	100,788
Gutenberg-DPO Jondurbin [2024b]	Creative Writing	Text	918
Imdb-dpo Yuasosnin [2024]	Summarization	Text	10,000
RewardBench Lambert et al. [2024]	Model Evaluation	Text	8,108
RewardBench2 Malik et al. [2025]	Model Evaluation	Text	1,865
VL-RewardBench Li et al. [2025b]	Model Evaluation	Multimodal	1,247
Stack Exchange Preference Lambert et al. [2023]	Instruction Tuning	Text	10,741,532
Nectar Zhu et al. [2023]	Preference Tuning	Text	182,954
OrcaSlimOrca Lian et al. [2023]	Instruction Tuning	Text	517,982
Alpaca Farm GPT-4 Dubois et al. [2024]	Instruction Tuning	Text	20,000
CodeSteer-DPO-Dataset Chen et al. [2025]	Code Generation	Text	4,462
ANAH Gu et al. [2025]	Hallucination Detection	Text	783
Unified Preference Dataset Wang et al. [2025b]	Preference Tuning	Multimodal	236,000
MM-IFDPO-23k Ding et al. [2025]	Instruction Tuning	Multimodal	22,555

Table 2: Overview of DPO datasets.

DPO algorithm aligns LLMs using datasets that consist of preference pairs. Typically, such a dataset includes a prompt x paired with two responses: a preferred (chosen) answer and a dispreferred (rejected) answer. Typically, human annotators are guided to evaluate and provide preferences among multiple responses generated from the same prompt. These datasets are denoted as human-labeled datasets in this survey. This process, however, is resource-intensive, especially when dealing with large volumes of data. To mitigate the high cost associated with human annotation, some datasets employ language models to synthesize preference pairs Cui et al. [2023a] Yu et al. [2024b] Zhang et al. [2024a]. These synthesized datasets are referred to as AI-labeled datasets in this survey. In this section, we will explore the two types of datasets in greater detail.

4.1 Human labeled

OpenAI:Summarize From Human Feedback. In downstream task like summarization, public online forum like Reddit can provide abundant text sources with a variety of topics. The TL;DR summarization dataset Völske et al.

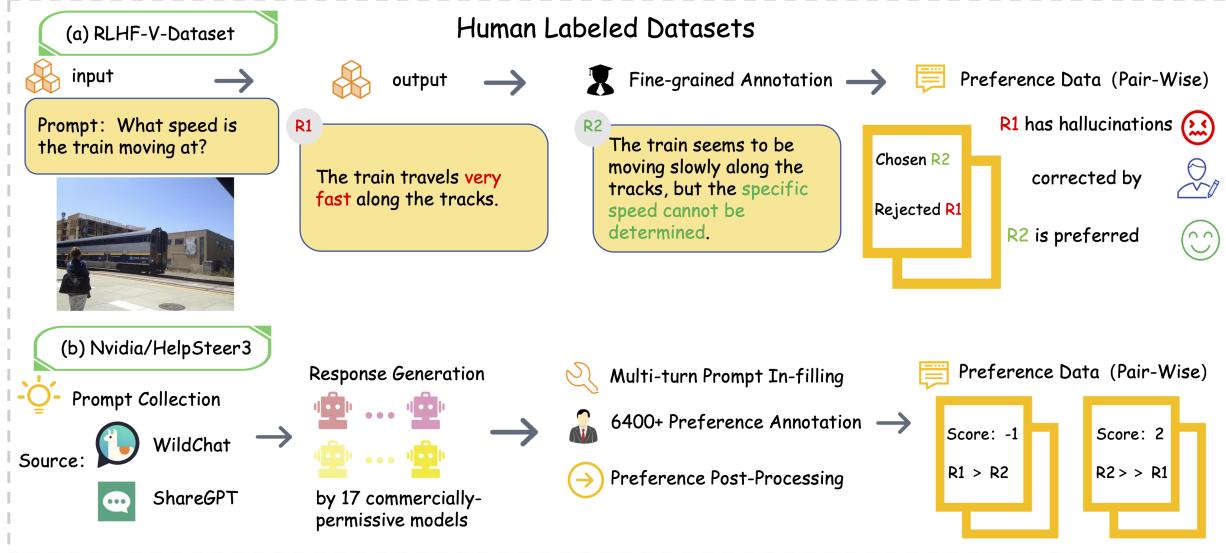


Figure 11: Two examples of Human labeled datasets: (a) RLHF-V-Dataset is a Pair-Wise preference dataset, where the unmodified large model responses are marked as rejected items, and the responses that humans have finely tuned are marked as chosen items. (b) HelpSteer3 is a Point-Wise preference dataset, with responses generated by 17 models and annotated by over 6,400 people. A negative score indicates that R1 is better, while a positive score indicates that R2 is better. The absolute value of the score reflects the degree of superiority or inferiority.

[2017] contains 3 millions post from reddit forum. To enhance the quality of data, Stiennon et al. [2022] filtered the dataset and left 123,169 posts with 5% as a validation set. Then Stiennon et al. [2022] collect human feedback on the filtered TL;DR dataset. To improve human labelers’s responses on the instruction, Stiennon team answered human labeler’s questions in a shared chat room, and provided regular feedback on their performance, which ensures high agreement on the judgements of responses. The new collected human preference dataset contains 64k text summarization examples including human-written responses and human-rated model responses, which could server as high quality sources of training data for DPO.

OpenAI:Webgpt comparisons. Webgpt comparisons Nakano et al. [2022] consist of 19,578 comparisons that are deemed appropriate for reward modeling, aiming at training a long-form question answering model aligned with human preferences. Each example in the dataset includes a pair of model-generated answers for a given question, along with associated metadata. Each answer is assigned a preference score from human annotators, allowing for the identification of the superior response. To gather these human preferences, Nakano et al. [2022] hire well-educated annotators, typically holding at least an undergraduate degree, due to the complexity of the tasks involved. Ultimately, this dataset captures human preferences in the generation of WebGPT, facilitating the optimization of the question-answering model to better align with human expectations.

Nvidia/HelpSteer. HelpSteer Wang et al. [2023c] is an open-source Helpfulness Dataset (CC-BY-4.0) designed to enhance model alignment for improved helpfulness, factual accuracy, and coherence, while allowing for adjustments in response complexity and verbosity. The dataset comprises 37,120 samples, each containing a prompt, a generated response, and five human-annotated attributes rated from 0 to 4, with higher scores indicating better performance. Prompts are sourced from various origins, with approximately half generated by Scale AI. The remaining prompts come from Open Question Answering, Generation, and Brainstorming tasks, as well as five specific tasks: Rewrite, Summarization, Classification, Extraction, and Closed Question Answering. Responses are produced by an early version of a 43 billion parameter in-house model, which generates up to four responses per prompt using sampling techniques to ensure diversity while maintaining reasonableness. The annotation process involved around 200 U.S.-based human annotators, who evaluated each response based on five attributes: Helpfulness, Correctness, Coherence, Complexity, and Verbosity, using a Likert-5 scale from 0 to 4.

Nvidia/HelpSteer2. HelpSteer2 Wang et al. [2024b], an open-source helpfulness dataset designed to train state-of-the-art reward models, comprises 21,362 samples, each containing a prompt, a response, and five human-annotated attributes, rated from 0 to 4, with higher scores indicating better performance. Consecutive samples (e.g., sample 1 with sample 2, sample 3 with sample 4) share the same prompt, allowing for preference pairs based on the helpfulness score (e.g., for training DPO or Preference RM), in addition to training SteerLM Regression RM. Over 95% of the prompts

are sourced from ShareGPT RyokoAI [2023], a platform where ChatGPT users voluntarily share their conversations, while a small proportion (5%) is generated by humans from Scale AI. Responses are produced by early versions of a mix of ten different in-house LLMs, generating two responses per prompt (one from each model) using sampling techniques to provide diverse yet reasonable answers.

Nvidia/HelpSteer3. HelpSteer3 Wang et al. [2025a], contains 40,476 preference samples, each containing a domain, language, context, two responses, an overall preference score between the responses, as well as individual preferences from up to 3 annotators. Each individual preference contains a preference score in addition to a concise reasoning for their preference in 1-2 sentences. 95% of the data is the training set and 5% is the validation set. The prompts are sourced from ShareGPT and WildChat-1M, covering four aspects: General tasks, Code, Multilingual, and STEM (Science, Technology, Engineering, and Mathematics). Responses were generated from a variety of 17 commercially-permissive models.

OpenAssistant/oasst1. OpenAssistant Conversations (OASST1) Köpf et al. [2023] is a human-generated, human-annotated assistant-style conversation corpus comprises 161,443 messages in 35 different languages, annotated with 461,292 quality ratings, resulting in over 10,000 fully annotated conversation trees. The corpus is the product of a worldwide crowd-sourcing initiative involving over 13,500 volunteers. The dataset consists of message trees, each starting with an initial prompt message as the root node, which can have multiple child messages as replies, and these child messages can themselves have additional replies. All messages have a role property, either "assistant" or "prompter," with roles alternating strictly between "prompter" and "assistant" from prompt to leaf node. Data collection utilized a web app interface, dividing the process into five separate steps: prompting, labeling prompts, adding reply messages as prompter or assistant, labeling replies, and ranking assistant replies. To ensure high quality in the collected human responses, the OpenAssistant team provides guidelines for all contributors. These guidelines clarify the meanings, scales, and criteria for assigning labels and rankings during the labeling and ranking tasks, and stipulate that assistant responses should be polite, helpful, concise, friendly, and safety-aware.

Stanfordnlp/SHP. SHP Ethayarajh et al. [2022] is a dataset consisting of 385K collective human preferences regarding responses to questions and instructions across 18 subject areas, ranging from cooking to legal advice. These preferences are designed to reflect the helpfulness of one response compared to another and are intended for training RLHF reward models. Each entry includes a Reddit post featuring a question or instruction and a pair of top-level comments, where one comment is more favored by users. The dataset leverages the principle that if comment A is posted after comment B yet receives a higher score, A is generally considered more preferred. However, if A precedes B, we cannot make the same inference, as its higher score may simply result from increased visibility. SHP emphasizes preference labels that indicate which response is more helpful, contrasting with previous work that often focused on identifying less harmful responses.

Tasksource/tasksource_dpo_pairs. The tasksource/tasksource_dpo_pairs dataset Sileo [2024] is specifically designed for use in DPO or RLHF, emphasizing expert-constructed examples over LLM-generated content. It encompasses data from various domains, particularly focusing on natural language inference (NLI) and logical reasoning tasks, to aid models in distinguishing between preferred and rejected responses. Structured in pairs, the dataset requires models to learn which of two responses is more favorable based on specific criteria. With over 4.8 million training examples, it also includes validation and test sets. Each example consists of a task description, a prompt, a chosen response, and a rejected response. This structure enables models to be fine-tuned to understand human-like preferences, enhancing their capability to generate useful and accurate responses in text generation tasks.

Openbmb/RLHF-V-Dataset. RLHF-V-Dataset Yu et al. [2024a] is a human-labeled dataset designed for the training of Multimodal Large Language Models (MLLMs). To address hallucination in MLLMs and enhance trustworthiness, Yu et al. [2024a] collect human-labeled fine-grained data for training and optimization, presenting RLHF-V. To be specific, RLHF-V collects high-quality fine-grained segment-level human corrections on the MLLMs responses with respect to question-answering and image description instructions. Human annotators are required to directly correct the hallucinated segments, which simultaneously introduces the preferred output. The final dataset contains a total of 5,733 preference pairs based on human annotation from 1.4k prompts from the instruction tuning dataset and image description from GPT-4. Using RLHF-V dataset in post-training effectively reduces hallucination in MLLMs.

Anthropic Helpful and Harmless dialogue dataset(HH-RLHF). HH-RLHFBai et al. [2022a] is a human-labeled public dataset that contains 169,352 rows of human preference pair of dialogues between a human and an automated assistant, including human preference data collected in Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback Bai et al. [2022a]. For helpfulness, human crowdworkers are required to chat with base models to engage in dialogue task like answering a question and discussing interest topics. Then, crowdworkers need to choose the more helpful and honest response between the two generated responses. For the harmlessness or red-teaming dataset, crowdworkers are given instructions to elicit harmful responses from their models (similar to helpfulness but to choose harmful responses). The overall Anthropic HH-RLHF dataset also contains Human-generated and annotated red

teaming dialogues from Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned Ganguli et al. [2022].

MM-RLHF. MM-RLHF Zhang et al. [2025] is a dataset containing 120k fine-grained, human-annotated preferences. It aims to align multimodal large language models with human preferences. This dataset was created by clustering and sampling from three areas (image understanding, video understanding, and multimodal safety), resulting in 30,000 high-quality queries. Additionally, MM-RLHF employs the most advanced response generation models to ensure response quality. Finally, after annotation, scoring, and ranking by over 50 professional annotators, the final dataset was obtained. Notably, this dataset provides the reasons for all scores and rankings, enhancing interpretability.

MMSafe-PO. Li et al. [2025a] MMSafe-PO comprises 5,667 multimodal instructions, each containing a text and a corresponding image. For each instruction, there is a chosen response and a rejected response. This dataset is transformed from the Anthropic-HH dataset, which recognizes the entities in text instructions and generates images of them, making this dataset applicable to the training of MLLMS. Like the Anthropic-HH dataset, it takes into account quantity, quality, and the harmless alignment objective.

4.2 AI labeled

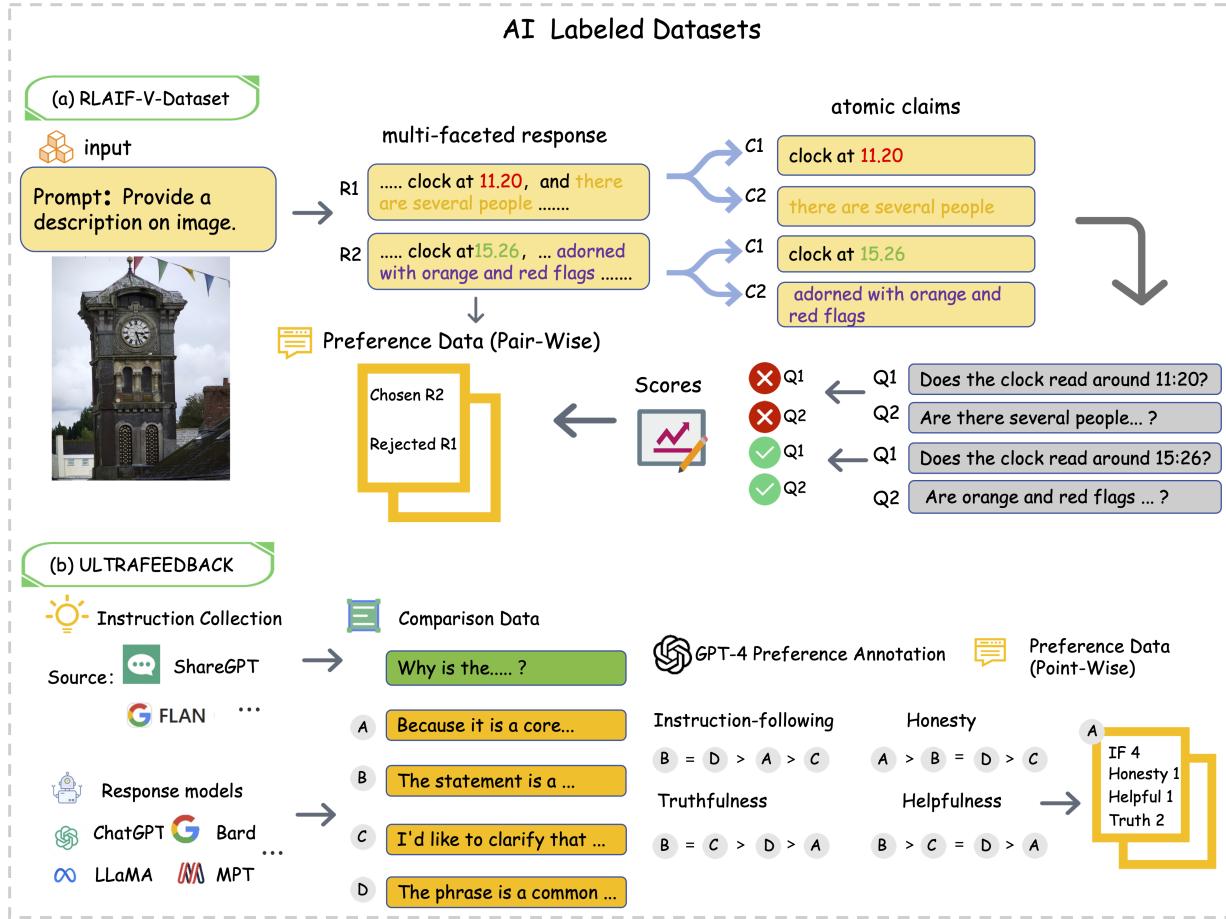


Figure 12: Two examples of AI labeled datasets: (a) RLAIF-V-Dataset is a Pair-Wise preference dataset. It designed a strategy to decompose responses, first breaking down responses into atomic claims and converting them into polar questions. This way, large models only need to reply with yes or no for score calculation, and ultimately form preference pairs based on the score's ranking. (b) ULTRAFEEDBACK is a Point-Wise preference dataset. It generates responses using multiple powerful large models and uses detailed prompts to have GPT-4 rank and score each group of responses on four dimensions to form the final preference set.

Openbmb/RLAIF-V-Dataset. RLAIF-V-Dataset Yu et al. [2024b] is a large-scale multimodal feedback dataset containing 83,132 high-quality preference pairs. The instructions are sourced from a diverse array of datasets, including

MSCOCO Lin et al. [2015], ShareGPT-4V Chen et al. [2023a], MovieNet Huang et al. [2020], Google Landmark v2 Weyand et al. [2020], VQA v2 Goyal et al. [2017], OKVQA Marino et al. [2019], and TextVQA Singh et al. [2019]. Additionally, the dataset incorporates image description prompts introduced in RLHF-V as long-form image-captioning instructions. Training on this dataset enables multi-modal models to achieve greater trustworthiness compared to both open-source and proprietary models.

Xinlai/Math-Step-DPO-10K. Math-Step-DPO-10K Lai et al. [2024] is a high-quality step-wise AI-labeled preference dataset with 10795 rows of preference pairs that are tailored for mathematical reasoning. The dataset construction involves three main steps. Each data sample includes four entries: 1) prompt x ; 2) initial reasoning steps; 3) preferred reasoning steps; and 4) undesirable reasoning steps. First, a mathematical problem prompt, accompanied by its ground-truth answer, is provided to a reference model, which generates a step-by-step solution using Chain-of-Thought (CoT). Next, solutions with steps differing from the ground-truth are selected to verify the correctness of each step, thereby identifying any erroneous steps. This process can be performed manually or with GPT-4. The undesirable reasoning steps are derived from these procedures. Multiple outputs are generated by inferring the reference model with the prompt x and the preceding reasoning steps to obtain the corresponding correct reasoning step for each sample. This procedure can be repeated multiple times until the ground-truth answer is achieved.

Openbmb/UltraFeedback. UltraFeedback Cui et al. [2023a] is a large-scale, fine-grained, and diverse preference dataset designed for training robust reward and critic models. Cui et al. [2023a] collected approximately 64,000 prompts from various sources, including UltraChat, ShareGPT, Evol-Instruct, TruthfulQA, FalseQA, and FLAN. These prompts were then used to query multiple LLMs, generating four different responses for each, resulting in a total of 256,000 samples. To ensure high-quality preference and textual feedback, they developed a fine-grained annotation instruction covering four aspects: instruction-following, truthfulness, honesty, and helpfulness. GPT-4 was employed to annotate the collected samples based on these guidelines. This dataset has been utilized by Saeidi et al. [2024] for the chat completion task, comprising 63,000 pairs of selected and rejected responses for specific inputs. The UltraFeedback-binarized dataset is used to train alignment models employing the DPO algorithm.

Argilla/Distilabel-capybara-dpo-7k-binarized. Capbara Argilla [2024] is a AI labelled dataset that contains 7536 conversations with chosen or rejected labels. Chosen and rejected pairs are formatted following OpenAI's conversation format with potentially several turns between a user and an assistant. Multi-turn dialogue data is key to fine-tune capable chat models. Multi-turn preference data has been used by the most relevant RLHF works (Anthropic). However, there are very few multi-turn open datasets for DPO/RLHF. This dataset is the first of a series of datasets to design for DPO/RLHF fine-tuning. Capybara generates three responses to the last user message using OSS 7B models and distilabel's LLMPool and the vLLM engine. From the 4 responses to each multi-turn dialogue, they use gpt-4-turbo to rank the quality of responses.

Malabonne/chatml-dpo-pairs. The dataset Malabonne [2024] is a pre-processed version of the Intel/orca-dpo-pairs dataset Mukherjee et al. [2023], formatted using ChatML. It consists of 12,000 examples from the Orca-style dataset, which includes pairs of prompts and responses. The dataset is designed for use with Direct Preference Optimization (DPO), which utilizes a pair of responses with chosen and rejected labels. In this dataset, the data format is a system prompt, a user query, and two responses: one from ChatGPT Bahrini et al. [2023] and one from Llama2-13b-chat Touvron et al. [2023b]. These responses are formatted, and the model is trained to prefer one response over the other, based on human or model feedback. The data has been processed to fit the ChatML format, making it suitable for various LLMs.

Magpie-Align/Magpie-Llama-3.1-Pro-DPO-100K-v0.1. The Magpie-Llama-3.1-Pro-DPO-100K Xu et al. [2024d] dataset contains 100,000 high-quality examples, making it suitable for fine-tuning LLMs using DPO, which involves comparing preferred and rejected responses to optimize the model's behavior based on human-like preferences. The Magpie-Llama-3.1-Pro-DPO-100K-v0.1 dataset was constructed as part of the Magpie Alignment project, which focuses on synthesizing high-quality alignment data for training large language models (LLMs) without relying on manual prompt engineering or seed questions. Instead, this dataset is generated by prompting pre-aligned LLMs, such as Meta's Llama 3.1 AI@Meta [2024], to autonomously produce instructional data.

Davanstrien/dataset-tldr-preference-dpo. The davanstrien/dataset-tldr-preference-dpo dataset davanstrien [2024] is designed for summarization and text generation tasks using DPO. It was constructed by creating concise summaries (tl;dr) Völske et al. [2017] of dataset cards, specifically aiming to highlight the most critical aspects of each dataset. These summaries were generated using LLMs like Meta-Llama-3-70B-Instruct AI@Meta [2024] and Nous-Hermes-2-Mixtral-8x7B-DPO "karan4d" "huemin_art" [2024]. The dataset includes prompts instructing the model to generate brief, informative summaries based on the dataset cards. The goal is to make these summaries useful for quickly understanding the dataset without unnecessary information, such as dataset size or license details. The dataset was filtered to ensure that only examples with meaningful differences between chosen and rejected responses were included.

Zerolink/Zsql-postgres-dpo. The zerolink/zsql-postgres-dpo dataset Zerolink [2024a] is designed for training models to convert natural language questions into SQL queries, specifically in the PostgreSQL dialect. This dataset contains around 200,000 pairs of DPO examples, making it ideal for text-to-SQL generation tasks. The construction of the dataset involves selecting SQL queries that are optimized for their syntactic and computational efficiency. For each example, the dataset includes a database schema, a natural language question, and two SQL queries: a "chosen" query (the preferred one) and a "rejected" query. These preferences are determined based on minimizing complexity and optimizing execution.

RewardBench Lambert et al. [2024] is a benchmark dataset evaluating reward model's performance on various queries. Rewardbench is a collection of chosen or reject responses as well as prompts over tasks such as chatting, reasoning, and safety. In chat task, Rewardbench collects subsets from AplacalEval Li et al. [2023b] including alpacaeval-easy, alpacaeval-length, alpacaeval-hard and MT Bench Zheng et al. [2023a] subset including, mt-bench-easy, mt-bench-medium. In safety category, Rewardbench collects subsets like efusals-dangerous, refusals-offensive, xtest-should-refuse Röttger et al. [2024], xtest-should-respond, do not answer Wang et al. [2023d]. In reasoning task, Rewardbench encompasses datasets like PRM Math Lightman et al. [2023], HumanEvalPack Muennighoff et al. [2024] on multiple programming languages. RewardBench also collects datasets from prior datasets including Anthropic Helpful, Stanford Human Preferences (SHP), and OpenAI's Learning to Summarize. The construction of dataset is done by scoring chosen-rejected-prompt with scalar scores. The success situation is when the chosen response attains higher scores than the reject responses. Rewardbench contains 8108 rows of samples in all.

RewardBench2 Malik et al. [2025], a new benchmark based on classification tasks, aims to provide challenging new data for RM evaluation based on accuracy. RewardBench2 uses new human prompts that have never been used for downstream evaluation (mostly from WildChat) as the main data source, promoting stricter evaluation and avoiding potential contamination with downstream evaluation goals. Its evaluation setting shifts from the common two-option (chosen and rejected) assessment to a four-option (one Chosen and three Rejected) assessment, reducing the random baseline from 50% to 25%, thus leaving more room for model improvement and enhancing score robustness, especially on more challenging subsets. RewardBench2 covers six domains, which are constructed through manual, programmatic, and LLM-based filtering techniques, aiming to improve the evaluation of challenging areas in existing RM benchmarks (Focus, Math, Safety) and add new challenging areas (Factuality, Precise Instruction Following, Ties).

VL-RewardBench Li et al. [2025b] is a comprehensive benchmark designed to evaluate the application of visual-language generation reward models (VL-GenRM) in visual perception, hallucination detection, and reasoning tasks. This benchmark consists of 1,247 high-quality examples, each carefully selected and composed of elements such as query, image, response, and human ranking. The queries are sourced from renowned datasets such as WildVision, RLAIF-V, Povid, and RLHF-V to ensure their quality. After screening by small models and human evaluation, challenging tasks were retained. Powerful commercial models generated responses and preference labels, and the final preference pairs were selected after human verification.

MMInstruction/VLFeedback. VLFeedback Li et al. [2023a] is a AI labeled dataset for large vision language models(LVLMs), aiming to enhance LVLMs' ability to generate factual and helpful responses following the visual context. Given multi-modal instructions from different datasets, 12 LVLMs are prompted to generate responses, promising that each data sample in VLFeedback contains 4 responses from different models. Then, GPT-4V is employed to evaluate the quality of the generated responses based on the criteria of helpfulness, visual factuality, and ethical considerations. At last, the LLM-annotated preference ranking dataset consists more than 380k comparison pairs and 80k multimodal instructions.

sqrti/SPA-VL. SPA-VL Zhang et al. [2024a] is a safety-specific dataset for safety alignment of Vision language Models(VLMs). SPA-VL is a GPT-4V labeled alignment dataset consisting of 100,788 samples in a wide range including 6 harmfulness domains, 13 categories, and 53 subcategories. Each preference pair contains the given question and image as well as the chosen response and rejected response. The primary goal of this dataset is to improve the harmlessness and helpfulness of VLMs while preserving their core capabilities. In practice, all the images in this dataset are collected from the LAION-5B dataset Schuhmann et al. [2022]. The accompany questions are generated by Gemini 1.0 Pro Vision Team et al. [2024]. To ensure the diversity of answers, SPA-VL employs 12 different models to sample different responses. Preference annotations are conducted by GPT-4V in order to determine which response is preferred and which one is dispreferred based on harmlessness and helpfulness.

Stack-exchange-preferences Lambert et al. [2023] is an open dataset containing questions and answers from Stack Overflow Data Dump. The purpose for this dataset is to enhance model's preference training and instruction fine-tuning. The questions from the stack overflow have been filtered to fit the criteria aligned with preference models.

Nectar Zhu et al. [2023] is the first 7-wise preference dataset consisting of 182,954 rows of samples generated through GPT-4. Nectar dataset is composed of diverse chat prompts, high-quality and diverse responses, and accurate ranking

labels. Prompts are sourced from many public dataset including lmsys-chat-1M Zheng et al. [2023b], ShareGPT, Antropic/hh-rlhf Bai et al. [2022a], UltraFeedback Cui et al. [2023a], Evol-Instruct Longpre et al. [2023], and Flan Xu et al. [2023]. Seven responses per prompt are primarily generated by a variety of models, including GPT-4, GPT-3.5-turbo, GPT-3.5-turbo-instruct.

Orca Lian et al. [2023] is subset of OpenOrca Mukherjee et al. [2023] dataset. OpenOrca collects augmented FLAN data collection Longpre et al. [2023], aiming at helping model to learning reasoning capability of the large foundation models through imitation learning. SlimOrca is a carefully curated subset of OpenOrca data, which applies an extra pass by using GPT-4 to remove answers that are considered as wrong based on human annotations from FLAN dataset. This step raises the overall quality of Orca dataset and reduces dataset size, minimizing the computational requirements.

Alpaca Farm GPT-4 Dubois et al. [2024] is a result dataset from the simulation framework for learning from human feedback. Dubois et al. [2024] come up with a new simulator that enables model to generate human preference for scientific research at low cost. Based on the previous instruction dataset, Alpaca farm is firstly designing GPT-4 prompt to generate preference similar to human annotators. Variability is also considered in the procedure of the pipeline for simulating human annotation. After the auto-evaluation, Alpaca farm can produce high-quality data that contain preference feedback similar to human annotation.

CodeSteer-DPO-Dataset Chen et al. [2025] This dataset is derived from the training requirements of the CodeSteer model. This model is responsible for guiding large language models to solve problems correctly and efficiently by using the text/code approach, with particular emphasis on its ability in symbolic calculation. This dataset is played by GPT-4o as both the instructor and the responder. The basic dataset is generated through multiple rounds of dialogue and the final preference dataset is obtained through preference ranking. This dataset contains 4,462 guidance comparison pairs. Each guidance contains guidance of different degrees on the problem. These guidance include guidance on aspects such as the selection of text/code, algorithm usage, and algorithm emphasis.

ANAH Gu et al. [2025] ANAH is a bilingual dataset that offers analytical annotation of hallucinations in LLMs within generative question answering. It can be used to train hallucination detector and perform factuality alignment. The dataset contains the literature corresponding to the topic. According to the questions selected from the literature, the answers from GPT3.5 and InternLM, as well as the analysis annotations for the responses of GPT3.5 and InternLM respectively. Each answer sentence in this dataset undergoes rigorous annotation, involving the retrieval of a reference fragment, the judgment of the hallucination type, and the correction of hallucinated content. ANAH consists of 12k sentence-level annotations for 4.3k LLM responses covering over 700 topics, constructed by a human-in-the-loop pipeline.

Unified Preference Dataset Wang et al. [2025b] A comprehensive human preference dataset that spans multiple vision-related tasks. It integrates existing datasets and preprocesses them to construct the first large-scale unified human preference dataset, which consists of approximately 236K data covering both image and video understanding and generation tasks. This dataset is a preprocessed and integrated version of eleven datasets, each of which contains preference rankings or ratings for images or videos.

MM-IFDPO-23k Construction Ding et al. [2025] This is a dataset used to enhance the Instruction Following ability of large models. The size of this dataset is 23k. Each row contains a conversation, an image, a selected answer and a rejected answer. It is used to help MLLMs better understand instructions and the responses users want. It defines 32 constraints to generate high-quality instruction-answer pairs. DPO training is conducted by generating rejection pairs by reducing the number of constraints.

5 Applications

With the rapid advancement of Large Language Models (LLMs) and Multi-modal Large Language Models (MLLMs), Direct Preference Optimization (DPO) has been effectively and robustly applied across various related fields. We simply categorize these applications into three distinct areas: LLMs, MLLMs, and a third category comprising other domains. This categorization allows us to highlight some representative works within each sector, showcasing the versatility and impact of DPO in model alignment.

5.1 Application on Large Language Models

Qwen2. At the fine-tuning stage, models within the Qwen2 Yang et al. [2024d] series undergo SFT as well as DPO. Through this preference learning, the models are endowed with enhanced abilities to follow instructions, improving their performance across diverse tasks and making them more responsive to human users. The preference learning of Qwen 2 models consists of two stages: offline training and online training Yang et al. [2024d]. In the online training

stage, the model further refines its performance by interacting with real-time feedback. During this phase, Qwen2 samples multiple candidate responses from the current policy, and a reward model evaluates and ranks them, selecting the most and least preferred responses to form preference pairs. These preference pairs are then fed back into the DPO process, allowing the model to iteratively improve its decision-making and instruction-following capabilities with each episode. This real-time refinement further aligns the model with human preferences, ensuring that it delivers more accurate and contextually appropriate responses. By integrating DPO at both stages, Qwen2 not only enhances its alignment with human preferences but also optimizes resource efficiency, making it suitable for handling large-scale models.

The Llama 3 Herd of Models. In Llama3 Herd of Models AI@Meta [2024], post-training procedure includes relatively simple methods such as SFT, reject sampling(RS) and DPO rather than complex reinforcement learning algorithms demanding extra complexity. After the supervised fine-tuning stage, several batches of preference data have been collected using the best performing models from the previous rounds of training. The distribution of the policy model is optimized in each training round based on the collected quality data. The Llama3 team also confirms that compared to on-policy algorithms like PPO Schulman et al. [2017b], DPO does reduce computational costs and perform better, especially on instruction following benchmarks like IFEval Zhou et al. [2023]. To leverage the advantages of DPO, Llama3 team applies two modifications to the original DPO algorithm. Firstly, special formatting tokens including header and terminator are masked out in from both chosen and rejected responses in the loss to stabilize the DPO training, which is based on the observation that these special tokens contributing to loss may induce undesired model behaviors such as tail repetition or abruptly generating termination tokens. Secondly, an additional negative log-likelihood (NLL) loss term is applied to DPO with a scaling coefficient of 0.2 on the chosen sequences, helping further stabilize DPO by maintaining desired formatting for generation and preventing the decrease of log probability of chosen responses Pang et al. [2024]. Enhanced Safety DPO approach is a crucial part of ensuring that the model not only follows instructions accurately but also adheres to strict safety guidelines, reducing the likelihood of generating harmful or inappropriate contents.

DeepSeek LLM. DeepSeek LLM DeepSeek-AI et al. [2024] base models are enhanced through post-training using DPO. DeepSeek-AI et al. [2024] observed an increase in repetitive output as the proportion of math SFT data increased, which they attributed to the repetitive reasoning patterns present in such data. Weaker models often struggle to generalize these patterns effectively, leading to undesirable repetitive responses. To address this issue and enhance the models' robustness, DPO was employed to further align the SFT models by utilizing a preference-based training approach. This approach leveraged pair-wise comparisons based on criteria such as helpfulness and harmlessness. Upon evaluation, the DPO-trained models demonstrated significant improvements across almost all key metrics, including performance in multilingual tasks such as creative writing, question answering, and instruction following. Moreover, the models also exhibited enhanced mathematical reasoning and logical deduction skills, underscoring the positive impact of the DPO training process in refining language model alignment.

Mixtral of Experts. Mixtral 8x7B Jiang et al. [2024a] is a Sparse Mixture of Experts (SMoE) model with publicly available weights. To improve its ability to follow complex instructions, Jiang et al. [2024a] applied a two-step training approach. First, they employed SFT on an instruction-focused dataset to enhance model's responsiveness to diverse commands. This was followed by DPO on a dataset with paired human feedback, allowing for further refinement based on user preferences. After completing the post-training process, Mixtral's performance significantly exceeded that of well-known models such as GPT-3.5 Turbo OpenAI [2022], Claude-2.1 Anthropic [2024], Gemini Pro Team et al. [2023], and Llama 2 70B Touvron et al. [2023c] in human evaluation benchmarks, highlighting the effectiveness of this combined training methodology.

Baichuan2. Baichuan 2 is a series of large-scale multilingual language models with 7 billion and 13 billion parameters, trained from scratch on a vast dataset of 2.6 trillion tokens Yang et al. [2023]. In the post-training stage, Baichuan2 is aligned with limited corpus of human-annotated data to improve the model's performance with respect to vulnerability safety issues. To be specific, Yang et al. [2023] is aligned with a red-teaming process that incorporates six types of safety attacks and over 100 granular safety value categories. Around 1K annotated data was used for initialization. By leveraging DPO, Baichuan 2 effectively addresses specific vulnerability issues, such as generating harmful or biased content, even with a limited amount of annotated safety-related data. By integrating DPO, Baichuan 2 demonstrates significant improvements in handling complex safety challenges, mitigating risks like the generation of misinformation, biased content, or other undesirable behaviors that can arise in large-scale multilingual models.

Yi: Open Foundation Models. Yi model family AI et al. [2024] encompasses a series of language and multi-modal models. The training of Yi model family proves that delicate fine-tuning framework can effectively save computational costs while maintaining same training results. In typical fine-tuning framework, reference or reward model is used to predict a batch of data or give a scalar reward value after certain action in specific environment. The time and computational resource spent on reference/reward model can not be neglected. In the next step, the target model will

use data from previous step to calculate loss and carry out gradient descent in order to update parameters. In the training framework of Yi model family, multi-model scheduling framework is adopted to facilitate support multiple backends for different LLMs in a single job AI et al. [2024]. Compared to PPO, DPO algorithm can facilitate the intermediate results from the reference mode. Reusing data from previous stage effectively improves the training speed and save the resources costs.

Besides the above general open-source large language models, LLMs also demonstrate their fabulous capability in multiple specific downstream tasks, including logical reasoning, instruction-following, anti-hallucination and code generation. Following subsections will detail DPO fine-tuning’s application on these domains and list related research.

5.1.1 Application on reasoning

Mathematical reasoning presents a significant challenge for large language models (LLMs) due to the need for high accuracy in long reasoning chains. While reinforcement learning methods like RLHF (using PPO) have been effective, they require resource-heavy training, including reward model development and policy optimization. Direct Preference Optimization (DPO) offers a simpler alternative, avoiding reward model training while achieving similar alignment objectives. However, DPO struggles with long-chain reasoning, such as in mathematical problems, as it lacks the granularity to identify specific errors within the reasoning process.

Step-DPO Lai et al. [2024] addresses this by introducing fine-grained supervision during fine-tuning. Instead of evaluating entire answers, Step-DPO treats individual reasoning steps as the units for optimization. This approach increases accuracy by focusing on identifying and correcting errors that typically arise mid-way through reasoning chains. Step-DPO maximizes the probability of selecting the correct next reasoning step while minimizing the probability of selecting incorrect steps, improving overall reasoning accuracy and factuality.

Step-Controlled DPO Lu et al. [2024b] also enhances models’ reasoning ability in step-wise granularity. To be specific, Step-Controlled DPO generates step-wise errors(dispreferred sample) in the concrete step among the long reasoning steps. By deploying DPO in the training stage, SCDPO enables model to understand the specific step that starts to be erroneous, supporting the model to generate right response.

In Iterative Reasoning Preference Optimization, Pang et al. [2024] introduce an iterative method that enhances LLMs’ reasoning, particularly focusing on CoT reasoning. This approach optimizes preferences among sampled CoT candidates by distinguishing between winning (correct) and losing (incorrect) reasoning steps. Compare to original DPO loss function, in this work, a variant of DPO is then trained, incorporating a negative log-likelihood (NLL) loss term for encouraging learning from winning pairs, which is essential for ultimate performance. The NLL loss is delicately designed improving winning responses. This method generates multiple responses in each iteration, constructs preference pairs based on correctness, and uses a modified DPO loss with an additional NLL term for training, alleviating errors in the reasoning steps.

DPO-augmented Self-Training(DPO-ST) Wang et al. [2024c] strengthens the original self-training algorithm to enhance models’ ability of CoT reasoning. Self-training is based on the prospective that models can learn from their own outputs and enhance predictive accuracy or decision-making capabilities. Incorporating direct preference optimization into traditional self-training pipeline can improve models’ chain-of-thought reasoning ability.

Large language models exhibit varying capabilities in mathematical reasoning between dominant languages, such as English, and non-dominant languages. This disparity can largely be attributed to the imbalance in pre-training and fine-tuning data across different languages Chen et al. [2023b]. To address this imbalance in reasoning abilities, MAPO She et al. [2024] introduces an innovative alignment method that leverages the robust reasoning capabilities of dominant languages to enhance the reasoning skills of non-dominant languages. By eliminating this requirement, She et al. [2024] effectively streamline the reasoning process, making it more accessible. By utilizing both DPO and iterative DPO, MAPO significantly enhances multilingual reasoning capabilities, ensuring that models are not only more effective in their performance across languages but also more equitable in their reasoning abilities. This advancement paves the way for more inclusive applications of language models in diverse linguistic contexts.

5.1.2 Application on instruction-following

The instruction-following ability of large language models refers to their capacity to understand, interpret, and execute commands given to them in natural language Lou et al. [2024b], Ouyang et al. [2022b]. To state-of-the-art large language models, the instruction-following capacity is essential as LLMs cannot leverage their enormous knowledge to perform various tasks and interpret users’ meaning without instruction-following fine-tuning stage. The instruction-following ability of LLMs is crucial for effectively executing commands in natural language. To enhance this capability, the AUTOIF Dong et al. [2024] framework introduces two approaches: SFT + Offline DPO and SFT + Iterative Online

DPO. For SFT + Offline DPO, after supervised fine-tuning, AUTOIF generates and filters various data scales, creating numerous positive and negative sample pairs for training. The AUTOIF framework employs both offline and iterative online DPO methods, significantly improving a model’s ability to understand and execute natural language instructions. Yang et al. [2024b] propose an innovative data augmentation technique that simplifies complex instructions by breaking them down into manageable sub-components. These components are then modified and reassembled, preserving the original context and complexity while adding variability essential for training and evaluating the instruction-following capabilities of large language models. Their method can indeed improve the instruction-following capacity of large language models by generating nuanced instruction variants while maintaining the original topic and context, which also allows models to better understand complex instructions. DPO enables LLMs to better finish the decomposed instructions with human preference.

LLMs struggle to generate responses that are aligned with given instructions, especially for complex instructions with multiple restraints. To handle the instruction following needs, Conifer Sun et al. [2024] introduces a new dataset dedicated to improving the instruction-following ability under multiple complex restraints that are closed to the real-world application situations. Conifer constructs a brand-new dataset with GPT-4’s insights on instructions with complex restraints, making it accessible for instruction-following training. The models are further trained by direct preference optimization(DPO). Conifer-7B-DPO is aligned with DPO algorithm, achieving the goal of enhancing instruction following ability of LLMs.

Large language models are prompt-sensitive as the quality of prompts greatly influences the instruction-following ability for LLMs’ performance on downstream tasks. The naive prompts induce the incorrect understanding of instructions. The free-form Free-form Instruction-oriented Prompt Optimization (FIPO; Lu et al. [2024c]) enables models to follow the core instruction within the naive prompts that are considered as model-agnostic manner by accurately optimization. DPO is employed in the process of prompt optimization based on the large-scale prompt preference dataset, achieving remarkable instruction-following ability.

With instruction-tuning, large language models learn to follow instructions to finish specific tasks. The ability or performance of LLMs can be evaluated on instruction following and corresponding responses. RS-DPO Khaki et al. [2024] further improves models’ instruction-following ability by combining rejection sampling and direct preference optimization. Evaluation on instruction-following benchmark demonstrates the effectiveness of the combination in enhancing models’ performance.

5.1.3 Application on anti-hallucination

The conventional alignment fails to enhance LLMs’ factual accuracy and leads to the generation of false fact (i.e. hallucination) Lin et al. [2024c]. Moreover, former study demonstrated that introducing new knowledge or unfamiliar texts in fine-tuning stage can encourage large language models to hallucinate Gekhman et al. [2024], Lin et al. [2024c], which makes SFT less factual as it trains on human labeled data that may be novel to LLM. Reinforcement learning also encourages LLM to hallucinate as the reward functions prefer longer responses on a diverse set of instructions that contain more details, which tends to stimulate the LLM to yield more false claims.

To tackle these factuality issues, alongside Retrieval-Augmented Generation (RAG; Gao et al. [2023], Zhao et al. [2024a]), Lin et al. [2024c] propose factuality-aware alignment (FLAME), which incorporates factuality-aware supervised fine-tuning (SFT) and reinforcement learning through direct preference optimization (DPO). DPO is central to this process, ensuring that LLM outputs align with factuality-aware objectives and enhance overall response quality. DPO, combined with factuality-aware alignment, offers a robust framework for improving the quality and factuality of LLM outputs.

To alleviate hallucinations in long-form generation, previous study Min et al. [2023] employs retrieval-augmented to detect hallucinations in models’ responses by decomposing long-form generation into atomic facts and verifying them by RAG. However, such retrieval-based algorithm can introduce irrelevant evidences. To better detect and judge hallucinations, Wang et al. [2024d] present a critic-based hallucination judge method. Since the quality of critiques are vital for judgement, HALU-J has been further fine-tuned with direct preference optimization. Experiment results demonstrate that DPO fine-tuning is helpful for performance and enhancing the quality of critiques.

Long form question answering often demands the large language model to generate response with more details, which also introduces hallucinations and factual inconsistencies. To handle this, Sachdeva et al. [2024] construct HaluQuestQA for hallucination detection and location dataset. Using HaluQuestQA dataset, Sachdeva et al. [2024] further propose a new prompt-based approach called error-informed refinement that refines the generated responses by learning feedback from the trained model. DPO-aligned refinement plays an crucial role in correcting errors in responses, reducing hallucinations in long form generation.

Tian et al. [2023] employ direct preference optimization to fine-tune language models for factuality. An estimated truthfulness score is computed to rank generated responses, resulting preferred response with higher truthfulness score. Through learning from generated preference rankings, DPO fine-tuning enhances models' factuality with/without retrieval-augmented generation.

5.1.4 Application on code-generation

Code language models(CLMs) are commonly evaluated by the passing rate to certain task or the accuracy to the generated code task. The running time for CLMs is not a standard criteria for code generation task. Code-Optimise Gee et al. [2024] evaluates the quality of generated by considering both correctness and runtime. Code-Optimise is optimized through direct preference optimization fine-tuning by learning from two dimensional signals including preference pairs over Quick versus Slow as well as Passed versus Failed. An improvement for performance can be observed by DPO fine-tuning.

Stable Code Pinnaparaju et al. [2024] is a serials of based code models that is aimed at code understanding, code interpretation, code completion, mathematical reasoning, and relevant software engineering tasks. Stable Code's programming capability is further reinforced through SFT and DPO. Training on the preference datasets such as UltraFeedback and HH-Anthropic dataset, Stable Code with 3B parameters outperform large language code models with 7B and 15B parameters, proving the effectiveness of DPO fine-tuning in post-training stage.

Code LLMs usually generate executable code responses to finish the required engineering tasks, assisting human engineers to complete coding domain instructions. However, code LLMs are struggling to generate fast codes for real-world performance because code LLMs only treat code snippets as texts. Nichols et al. [2024] introduce a new alignment to enhance code LLMs' performance for generating fast code. Using the variant of DPO fine-tuning, Direct Performance Alignment (DPA; Nichols et al. [2024]) fine-tunes code LLMs by training on faster and slower code pairs for preference optimization, achieving new benchmark for code performance.

Large language models have demonstrated comprehensive ability for various downstream tasks including code generation. General LLMs fails to accomplish all code generation tasks especially on competition level programming tasks as these tasks include many corner cases that are not trained in the pre-training stage. Instruct-Code-Llama Liu et al. [2024e] introduces the new fine-tuning approach using Reinforcement Learning with Online Judging Feedback. In RL stage, Instruction-Code-Llama is further aligned using DPO with high-quality data, which achieves less complexity in time and space of generated codes and strengthens the efficiency. The above subsections elaborate the current application of DPO fine-tuning in downstream tasks for forefront LLMs research. Actually, DPO is also employed in evaluation task. RewardBench Lambert et al. [2024] constructs a new benchmark for reward models to evaluate reward models' capabilities in multiple domains such as chat, reasoning, safety as well as helpfulness and harmlessness. Direct judgement preference optimization Wang et al. [2024e] deploys the DPO fine-tuning to enable the model to learn from the positive and negative samples, enhancing models' evaluation capabilities across different domain of tasks including reasoning, judgement, and models' response-comprehension.

5.2 Application on Multi-modal Understanding and Generation

DPO and its improved variants have been widely applied to image, video, and speech processing, among other areas. We categorize these applications into two major parts: multi-modal understanding (e.g., image captioning, video captioning, etc.) and multi-modal generation (e.g., image synthesis, video generation, etc.). This includes Multi-modal Large Language Models (MLLMs) and Diffusion models. We introduce several key applications and improvements of DPO in these domains, showcasing its impact on model alignment.

5.2.1 Multi-modal Understanding

Image understanding: A prevalent issue in Large Vision Language Models (LVLMs) is hallucination, where the generated responses fail to align with the provided image and text contexts, significantly restricting the usages of LVLMs Liu et al. [2024f], Yin et al. [2023], Cui et al. [2023b]. Generally, humans prefer non-hallucinatory responses over hallucinatory ones. Consequently, some studies have reinterpreted the problem of model hallucinations as a preference optimization task, where DPO is widely employed.

Li et al. [2023a] collected large-scale of AI feedback from various models and directly applied DPO on LVLMs for distillation from AI feedback. Zhou et al. [2024b] curated preference dataset by constructing dispreferred responses with using distorted images as input to trigger hallucinations or injecting hallucinations into the policy model's responses.

Yu et al. [2024a,b] proposed a token-level variant of DPO to mitigate hallucinations in dense captioning tasks. They found that some key words in the chosen responses required more intensive reinforcement. Specifically, they gathered

human preferences Yu et al. [2024a] or AI feedback Yu et al. [2024b] in the form of segment-level corrections for hallucinations and implemented a dense variant of DPO. Xiao et al. [2024] proposed a detect-then-rewrite pipeline for differentiating and mitigating hallucinations, and employed an variant of DPO which adaptively assigns each rejected response a hallucination severity score.

In Llama 3 Herd of Models AI@Meta [2024], authors adopted DPO in post-training stage and have find that instead of always freezing the reference model, updating it in an exponential moving average (EMA) fashion every k-steps helps the model learn more from the data, resulting in better performance in human evaluations. Besides, they observed that the vision DPO model consistently performs better than its SFT starting point in human evaluations for every finetuning iteration.

Video understanding: Zhang et al. [2024b] introduced a novel framework named LLaVA-Hound that utilizes detailed video captions as a proxy of video content. Specifically, they extracted dense captions on video and then employed a large language model with these captions to score video QA. Their empirical results demonstrate this method's robust alignment with the reward from GPT-4V OpenAI [2023a], which can take images (i.e. video frames) as input. Additionally, Zhang et al. [2024d] developed LLaVA-Next-Video by supervised fine-tuning (SFT) LLaVA-Next-Image on video data, achieves better video understanding capabilities compared to LLaVA-Next-Image. Further, they employed LLaVA-Hound Zhang et al. [2024b] which aligns the model response with AI feedback using DPO, showing significant performance boost.

Audio understanding: In Qwen2-Audio's development Chu et al. [2024], the authors directly employed vanilla DPO to align the SFT model with factuality and human preference. Ye et al. [2024] developed a MLLM named CAT that responds to audio-visual content, and they have found that although MLLMs are equipped with multimodal understanding, they cannot perceive ambiguity. Therefore, they have reinterpreted ambiguity mitigation as a model preference optimization task and proposed an AI-assisted Ambiguity-aware Direct Preference Optimization (ADPO) method. Specifically, they took ambiguous responses that express the lack of clarity of specific audio-visual objects as negative responses, and then they utilized GPT to rewrite them into positive responses. After the multimodal training, they performed ADPO to align the model to generate the positive response (i.e. the precise description after the rewrite) and reject the negative response (i.e. the ambiguous description).

5.2.2 Multi-modal Generation

Image generation: Aligning text-to-image (T2I) diffusion model with preference has been gaining increasing research attention. Some recent works have applied DPO on diffusion models to better align synthetic images with human preferences (e.g. aesthetic, fidelity, NSFW, et etc).

Wallace et al. [2023]re-formulated DPO to account for a diffusion model notion of likelihood, utilizing the evidence lower bound to derive a differentiable objective. In concurrent works, Yang et al. [2024c] reinterpreted the denoising process inherent in diffusion models as a multi-step MDP. Then they extend the theory of DPO to MDP, which allows us to apply the principles to effectively translate human preferences into policy improvements in diffusion models.

Previous works, such as Wallace et al. [2023], Yang et al. [2024c], have explored the align T2I generation using DPO. These approaches typically works under the bandit assumption, focusing on a latent reward applied to the entire diffusion denoise process. However, this ignores the sequential nature of the generation process. To address these limitations, Yang et al. [2024e] propose a fine-grained approach with a dense reward. They develop a tractable objective focusing on the initial stages of the denoising chain. Besides, their method incorporates temporal discounting into implicit reward objectives, which better conform to the hierarchical structure of T2I generation.

Audio generation: Majumder et al. [2024] leveraged an existing text-to-audio model, Tango, to synthetically create a preference dataset. In this dataset, each prompt is associated with a chosen audio output and several rejected audio outputs, which the diffusion model learns from. The rejected outputs theoretically contain some concepts from the prompt either missing or ordered incorrectly. The authors fine-tuned the text-to-audio model Tango with diffusion-DPO Wallace et al. [2023] on the preference dataset. Their results demonstrate that this approach yields improved audio outputs over Tango, as evidenced by both automatic and manual evaluation metrics.

Current emotional text-to-speech (TTS) models Diatlova and Shutov [2023], Li et al. [2024b] only learn the correct emotional outputs without fully comprehending other emotion characteristics, limiting their ability to capture the nuances between different emotions. Gao et al. [2024] propose a controllable Emo-DPO approach, which employs direct preference optimization to differentiate subtle emotional nuances between emotions through optimizing towards preferred emotions over less preferred emotional ones. Zhang et al. [2024c] have employed DPO in speech generation to align speech generation to human preferences, and they have found that DPO contributes to improvement in speech generation and validated that iterative DPO can consistently enhance the generated speech's quality.

Cho et al. [2024] have found that current instruction-tuned LLMs are primarily trained with textual data, which often results in a misalignment with the specific needs of other modalities including speech. To address this issue, they gathered a novel dataset with 20,000 speech-based preference samples. These samples were created using a wide range of prompts to capture different aspects of speech suitability and were labeled by annotators who listened to pairs of responses. They then applied DPO to some instruction-tuned LLMs to enhance the speech suitability of the generated responses.

5.3 More Applications

Widatalla et al. [2024] align a structure-conditioned language model to generate stable protein sequences by encouraging the model to prefer stabilizing over destabilizing variants given a protein backbone structure. Empirical results indicate that ProteinDPO has learned generalizable information from its biophysical alignment data.

Generative pre-trained transformer is deployed to address complex issue in pharmaceutical research, especially in drug delivery system. Hu et al. [2024] incorporate reinforcement learning from human feedback Strategy for Photoresponsive Molecules in Drug Delivery. which is based on base model to fine-tune for addressing smart drug administration. Particularly, in the chemical knowledge datasets with different molecules, base model is learning from the dictionary object by accepting the recommended molecules and rejecting the dispreferred samples. Experimental results demonstrate the effectiveness of DPO algorithm in this domain.

ConPro Nguyen et al. [2024] leverages direct preference optimization in helping model to learn severity representation in medical images. Ranking degree of severity in recommendation system requires the learning from preference comparison between various severity classes and normal classes. ConPro achieves new benchmark for classification tasks in severity ranking domain, demonstrating the potential of learning from comparison by DPO algorithm.

Preference learning from human feedback can also be applied to explore text-to-motion generation. Sheng et al. [2024] fine-tunes MotionGPT Jiang et al. [2024b] by direct preference optimization with collected preference pairs generated by MotionGPT, which guides the based model to choose the sample with higher policy-assigned probability suggested by the reference policy. Experimental results demonstrate that in this work, DPO algorithm outperforms RLHF in alignment metrics, endowing a more effective approach in text-to-motion tasks.

Softmax-DPO Chen et al. [2024c] utilizes the DPO’s characteristic of learning from both positive samples and negative samples to instill ranking information to LLMs for the sake of personalized preference recommendation. Learning from the multiple negative samples, Softmax-DPO leverages language models’ reasoning capabilities to predict the next user-preferred item rather than the next-token probability, boosting preference recommendation based on LLMs with direct preference optimization.

DMPO Bai et al. [2024] aims to bridge the gap between the general data used in pre-training stage and the specific recommendation tasks in application. Different from original DPO fine-tuning’s intention of maximizing the positive sample, DMPO intends to enhance models’ performance on cross-domain recommendations by not simply maximizing the probability of positive samples but also minimizing probability of multiple negative samples. Experimental results indicate that DMPO approach is enhancing models’ performance in recommendation and outperforms traditional sequential approaches as well as LLM-based approaches.

6 Discussion and Open Challenges

Numerous analytical studies and variants of DPO for large language and multi-modal models have been proposed, as detailed above. However, new challenges related to DPO continue to emerge. Therefore, several open issues remain that warrant thorough discussion and investigation. In this section, we provide detailed suggestions for future research directions, including preference dataset construction, generalization of DPO and more complex applications.

Preference Feedback. Typically, the preference pairs used in DPO consist of a prompt paired with a preferred (chosen) response and a dispreferred (rejected) response, either labeled manually or automatically. However, current methods for constructing preference datasets have certain limitations that need to be addressed in future research.

- First, current automatic methods for dataset construction mostly use proprietary AI models, such as GPT-4, to replace human as preference annotator Lee et al. [2023]. Therefore, the annotation quality is restricted by the capability of AI models, which may lead to inferior alignment preference. Furthermore, as the intelligence of AI models continues to advance, it will become increasingly difficult for even human annotators to accurately distinguish between preferred and dispreferred answers Burns et al. [2023], Zhao et al. [2024b]. Therefore,

how to develop an accurate and efficient method to provide large-scale human oversight in the future calls for significant attention.

- Second, current preference feedback primarily comes from either humans or AI models, relying on a single source and failing to leverage multi-source feedback from real-world environments, tools and specialized models Liu et al. [2024e], McAleese et al. [2024]. Additionally, how to effectively harness such heterogeneous feedback to improve the helpfulness, harmlessness and honesty of target models remains an important area for future research.
- Third, most preference feedback is *coarse-grained* (i.e., at the instance level) with some recent efforts exploring fine-grained AI feedback Yu et al. [2024a], Wu et al. [2024d], Xiao et al. [2024], Lai et al. [2024], Lu et al. [2024b], such as feedback at the sentence or token level. However, this fine-grained feedback has primarily been gathered for specialized tasks like mathematical reasoning or visual question answering, which is insufficient for the development of a general model like OpenAI o1 OpenAI [2024b] that can generate long internal chain-of-thoughts. Thus, future research is required to develop accurate and efficient methods for providing such fine-grained feedback.

Generalization of DPO. Owing to its stable training, competitive performance, and computational efficiency, DPO has been widely adopted as an alternative in various applications Rafailov et al. [2023, 2024a], Ethayarajh et al. [2024], Saeidi et al. [2024]. However, DPO also has been shown to exhibit inferior generalization compared to online methods Xu et al. [2024a], Ivison et al. [2024]. To enhance its generalization ability, we propose the following potential directions for future research.

- First, from the data perspective, existing studies suggest that the inferior performance of DPO stems from the using of offline data Xu et al. [2024a], Ivison et al. [2024], Wang et al. [2024a]. As a result, subsequent studies have introduced various DPO variants to incorporate online data from the policy model into alignment learning Yuan et al. [2024d], Xiong et al. [2024], Pang et al. [2024], with the aim of improving the generalization ability of DPO. However, the performance of current methods, such as self-rewarding Yuan et al. [2024d] and iterative DPO Pang et al. [2024], plateaus after several iterations. Thus, how to further enhance the ability of DPO to exploit additional online data remains a open question.
- Second, from the perspective of learning objective, DPO align models with human preferences via a preferred answer and a dispreferred answer. However, the degree of preference between different preference pairs is different, which is ignored by the original learning objective of DPO. Consequently, various studies have explored methods to enhance the learning objective’s awareness of the preference difference between the rewards of chosen and rejected responses Meng et al. [2024], Xu et al. [2024b], Xiao et al. [2024], Azar et al. [2023]. However, some issues remain underexplored. For example, enabling alignment awareness of nuanced differences in preference pairs is crucial in scenarios where each token carries significant importance. Additionally, teaching models to reason through DPO alignment poses the challenge of fostering an awareness of intrinsic reasoning abilities in chosen responses, rather than merely following reasoning formats or specific tokens.
- Lastly, both explicit and implicit reward models predict human preferences using static internal parameters, which inherently impose limitations, such as the inability to access real-time information and the lack of vertical domain knowledge. Consequently, tool-augmented or knowledge-augmented reward modeling has been investigated to construct a more accurate and comprehensive reward signal Li et al. [2024c], Tian et al. [2023], Lin et al. [2024c]. Incorporating these methods into implicit reward modeling, however, remains a promising yet challenging area of research.

More Applications. Although DPO has been widely adopted in various applications, we believe it has the potential to be applied to a broader range of applications.

- Mixed-modal models Lu et al. [2023], Team [2024], He et al. [2024], Xie et al. [2024] aimed to develop a single unified model capable of handling a wide variety of tasks across vision, language-vision, language, and potentially additional modalities. The core idea behind these models is to homogenize each type of input and output into sequences of discrete vocabulary tokens, thereby unifying multi-modal understanding and generation. DPO interprets autoregressive token generation as a MDP, which makes it inherently suitable for aligning mixed-modal models.
- The reasoning capabilities of LLMs could be further enhanced through DPO algorithm. The emergence of OpenAI o1 model signifies the new era of shifting the computational resources further into post-training OpenAI [2024b]. Hiding chain of thought empowers o1 remarkable reasoning capabilities, which hints that the

pivot of LLM training can be moved to post-training especially reinforcement learning stage after the model reaches a certain scale in pre-training. As a lightweight alignment, DPO algorithm is employed to simplify the complicated multi-model training in the post-training stage while maintaining quality performance. Variant of DPO objective may guide the training model to be rewarded for the more fine-grained chain of thought, reinforcing the learning of correct CoT over the multiple incorrect reasoning steps. The potential of DPO algorithm in enhancing reasoning has not been fully explored in current research.

- Current works in video generation seek to achieve real-world simulations Li et al. [2024d]. Despite these efforts, challenges persist, such as violations of physical laws and issues related to controllability, including NSFW content and privacy concerns. Recent works by Prabhudesai et al. [2024] explore backpropagating gradients from trained reward models to a video diffusion model. The direct application of DPO to video generation models remains largely unexplored. Given that DPO is successfully utilized for safety alignment and hallucination mitigation, its integration into video generation models offers promising potential.

7 Conclusion

In this work, we provide a comprehensive review of recent advancements in DPO, a widely used lightweight preference learning method, covering aspects such as preference feedback, theoretical analyses, variants, and applications. Additionally, we propose several future research directions, with the aim of offering insights to the research community on aligning foundational models, including LLMs, MLLMs, and beyond.

References

- OpenAI. Introducing chatgpt. 2022.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- OpenAI. Gpt-4 technical report, 2024a. URL <https://arxiv.org/abs/2303.08774>.
- 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.
- OpenAI. Gpt-4v(ision) system card. 2023a.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.
- Gemini Team, 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.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.
- Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018. URL <https://arxiv.org/abs/1811.07871>.
- 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 arXiv:2112.00861*, 2021.
- OpenAI. Introducing superalignment. <https://openai.com/blog/introducing-superalignment>, 2023b. Accessed on July 5, 2023.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022a.
- Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=TyFrP0KYXw>.

- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. URL <https://arxiv.org/abs/2309.14525>.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017a.
- Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémie Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitri Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2023, 2023. URL <https://openreview.net/forum?id=bx24KpJ4Eb>.
- Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL <https://openreview.net/forum?id=JYtwGwIL7ye>.
- Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D. Dragan, and Daniel S. Brown. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=R0Xxvr_X3ZA.
- 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, 2023. URL <https://arxiv.org/abs/2304.05302>.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment, 2023. URL <https://arxiv.org/abs/2304.06767>.
- Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization, 2024b. URL <https://arxiv.org/abs/2309.06657>.
- Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment, 2024. URL <https://arxiv.org/abs/2306.17492>.
- 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, 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, 1952a.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024.
- Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback, 2024. URL <https://arxiv.org/abs/2406.09279>.
- Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO superior to PPO for LLM alignment? A comprehensive study. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024a. URL <https://openreview.net/forum?id=6XH8R7YrSk>.
- Amir Saeidi, Shivanshu Verma, and Chitta Baral. Insights into alignment: Evaluating dpo and its variants across multiple tasks, 2024. URL <https://arxiv.org/abs/2404.14723>.
- Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of rlhf, 2024a. URL <https://arxiv.org/abs/2309.06256>.
- Keming Lu, Bowen Yu, Fei Huang, Yang Fan, Runji Lin, and Chang Zhou. Online merging optimizers for boosting rewards and mitigating tax in alignment. *arXiv preprint arXiv:2405.17931*, 2024a.
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023.
- Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation, 2024b.

- Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model, 2024. URL <https://arxiv.org/abs/2403.07691>.
- Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024.
- Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning, 2024b. URL <https://arxiv.org/abs/2407.00782>.
- Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhenlun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback, 2024. URL <https://arxiv.org/abs/2404.14233>.
- Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization, 2024. URL <https://arxiv.org/abs/2404.11999>.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.
- Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023a.
- Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.
- Yong Lin, Skyler Seto, Maartje ter Hoeve, Katherine Metcalf, Barry-John Theobald, Xuan Wang, Yizhe Zhang, Chen Huang, and Tong Zhang. On the limited generalization capability of the implicit reward model induced by direct preference optimization, 2024b. URL <https://arxiv.org/abs/2409.03650>.
- Ziniu Li, Tian Xu, and Yang Yu. Policy optimization in rlhf: The impact of out-of-preference data, 2024a. URL <https://arxiv.org/abs/2312.10584>.
- Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms, 2024a. URL <https://arxiv.org/abs/2406.10216>.
- Chen Jia. Generalizing reward modeling for out-of-distribution preference learning. In Albert Bifet, Jesse Davis, Tomas Krilavičius, Meelis Kull, Eirini Ntoutsi, and Indrė Žliobaitė, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 107–124, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-70362-1.
- Yixin Liu, Pengfei Liu, and Arman Cohan. Understanding reference policies in direct preference optimization, 2024c. URL <https://arxiv.org/abs/2407.13709>.
- Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of dpo: A theoretical perspective, 2024. URL <https://arxiv.org/abs/2404.04626>.
- Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^* : Your language model is secretly a q-function, 2024a. URL <https://arxiv.org/abs/2404.12358>.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models, 2024a. URL <https://arxiv.org/abs/2401.01335>.
- Yunhao Tang, Daniel Zhaoan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. Understanding the performance gap between online and offline alignment algorithms, 2024. URL <https://arxiv.org/abs/2405.08448>.
- Shiqi Wang, Zhengze Zhang, Rui Zhao, Fei Tan, and Cam Tu Nguyen. Offline rlhf methods need more accurate supervision signals, 2024a. URL <https://arxiv.org/abs/2408.09385>.
- Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Iterative preference optimization with the pairwise cringe loss, 2024c. URL <https://arxiv.org/abs/2312.16682>.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online ai feedback, 2024a. URL <https://arxiv.org/abs/2402.04792>.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024a.
- Lichang Chen, Juhai Chen, Chenxi Liu, John Kirchenbauer, Davit Soselia, Chen Zhu, Tom Goldstein, Tianyi Zhou, and Heng Huang. Optune: Efficient online preference tuning, 2024b. URL <https://arxiv.org/abs/2406.07657>.
- 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, 2024. URL <https://arxiv.org/abs/2305.14387>.

- Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf, 2023. URL <https://arxiv.org/abs/2310.03716>.
- Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. Is stack overflow obsolete? an empirical study of the characteristics of chatgpt answers to stack overflow questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24*, page 1–17. ACM, May 2024. doi: 10.1145/3613904.3642596. URL <http://dx.doi.org/10.1145/3613904.3642596>.
- Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. How far can camels go? exploring the state of instruction tuning on open resources, 2023b. URL <https://arxiv.org/abs/2306.04751>.
- Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 4998–5017, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.findings-acl.297>.
- Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. Following length constraints in instructions, 2024b. URL <https://arxiv.org/abs/2406.17744>.
- Wei Liu, Yang Bai, Chengcheng Han, Rongxiang Weng, Jun Xu, Xuezhi Cao, Jingang Wang, and Xunliang Cai. Length desensitization in directed preference optimization, 2024d. URL <https://arxiv.org/abs/2409.06411>.
- Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. Spo: Multi-dimensional preference sequential alignment with implicit reward modeling, 2024a. URL <https://arxiv.org/abs/2405.12739>.
- Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Jie Xin Wang, Huimin Chen, Bowen Sun, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable preference optimization: Toward controllable multi-objective alignment, 2024b. URL <https://arxiv.org/abs/2402.19085>.
- OpenAI. Learning to reason with llms. Sep 2024b. URL <https://openai.com/index/learning-to-reason-with-llms/>.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. doi: 10.48550/arXiv.2405.09818. URL <https://github.com/facebookresearch/chameleon>.
- Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=SQnitDuow6>.
- Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing llm reasoning generalists with preference trees, 2024c.
- Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim, and Chanjun Park. sdpo: Don't use your data all at once, 2024.
- Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms, 2024. URL <https://arxiv.org/abs/2406.18629>.
- Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khaltman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, Chi Jin, Tong Zhang, and Tianqi Liu. Building math agents with multi-turn iterative preference learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=WjKea8bGFF>.
- Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization, 2024. URL <https://arxiv.org/abs/2404.19733>.
- Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. β -dpo: Direct preference optimization with dynamic β , 2024a. URL <https://arxiv.org/abs/2407.08639>.
- Sangkyu Lee, Janghoon Han, Hosung Song, Stanley Jungkyu Choi, Honglak Lee, and Youngjae Yu. Kl penalty control via perturbation for direct preference optimization. *arXiv preprint arXiv:2502.13177*, 2025.
- Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=H0qIWXXLUR>.
- Biqing Qi, Pengfei Li, Fangyuan Li, Junqi Gao, Kaiyan Zhang, and Bowen Zhou. Online dpo: Online direct preference optimization with fast-slow chasing, 2024. URL <https://arxiv.org/abs/2406.05534>.

- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2024d. URL <https://arxiv.org/abs/2401.10020>.
- Avinandan Bose, Zhihan Xiong, Aadirupa Saha, Simon Shaolei Du, and Maryam Fazel. Hybrid preference optimization for alignment: Provably faster convergence rates by combining offline preferences with online exploration. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025. URL <https://openreview.net/forum?id=YBBXsVta2x>.
- Paria Rashidinejad and Yuandong Tian. Sail into the headwind: Alignment via robust rewards and dynamic labels against reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=I8af9JdQTy>.
- Sungdong Kim and Minjoon Seo. Rethinking the role of proxy rewards in language model alignment, 2024. URL <https://arxiv.org/abs/2402.03469>.
- Shiva Kumar Pentyala, Zhichao Wang, Bin Bi, Kiran Ramnath, Xiang-Bo Mao, Regunathan Radhakrishnan, Sitaram Asur, Na, and Cheng. Paft: A parallel training paradigm for effective llm fine-tuning, 2024. URL <https://arxiv.org/abs/2406.17923>.
- Tingchen Fu, Deng Cai, Lemao Liu, Shuming Shi, and Rui Yan. Disperse-then-merge: Pushing the limits of instruction tuning via alignment tax reduction, 2024. URL <https://arxiv.org/abs/2405.13432>.
- Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10586–10613, 2024a.
- Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J Su, and Li Shen. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=51tMpvPNSm>.
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022.
- Damien Sileo. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1361>.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL <https://arxiv.org/abs/2112.09332>.
- Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. Helpsteer: Multi-attribute helpfulness dataset for steerlm, 2023c. URL <https://arxiv.org/abs/2311.09528>.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023. URL <https://arxiv.org/abs/2304.07327>.
- Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with \mathcal{V} -usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/ethayarajh22a.html>.
- Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024b. URL <https://arxiv.org/abs/2406.08673>.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024a.

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL <https://arxiv.org/abs/2204.05862>.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023a. URL <https://arxiv.org/abs/2306.05685>.
- Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages, 2025a. URL <https://arxiv.org/abs/2505.11475>.
- Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Fan Yang, Zhang Zhang, Tingting Gao, Di Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mm-rlhf: The next step forward in multimodal llm alignment, 2025. URL <https://arxiv.org/abs/2502.10391>.
- Yongqi Li, Lu Yang, Jian Wang, Runyang You, Wenjie Li, and Liqiang Nie. Towards harmless multimodal assistants with blind preference optimization, 2025a. URL <https://arxiv.org/abs/2503.14189>.
- Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness, 2024b. URL <https://arxiv.org/abs/2405.17220>.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023a.
- Argilla. Distilabel capybara. <https://huggingface.co/datasets/argilla/distilabel-capybara-dpo-7k-binarized>, 2024.
- Malabonne. Mlabonne/chatmldpopairs. <https://huggingface.co/datasets/mlabonne>, 2024.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024. URL <https://arxiv.org/abs/2403.13787>.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing, 2024d. URL <https://arxiv.org/abs/2406.08464>.
- Zerolink. zerolink/zsql-postgres-dpo. <https://huggingface.co/datasets/zerolink/zsql-postgres-dpo>, 2024a.
- Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7b: Improving llm helpfulness & harmlessness with rlaif, November 2023.
- Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichek Vong, and "Teknium". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023. URL <https://huggingface.co/Open-Orca/SlimOrca>.
- Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, and Jing Shao. Spa-vl: A comprehensive safety preference alignment dataset for vision language model, 2024a. URL <https://arxiv.org/abs/2406.12030>.
- Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023a.
- Jondurbin. Jondurbin/truthy-dpo-v0.1. <https://huggingface.co/datasets/jondurbin/py-dpo-v0.1>, 2024a.
- Zerolink. Jondurbin/truthy-dpo-v0.1. <https://huggingface.co/datasets/jondurbin/truthy-dpo-v0.1>, 2024b.
- Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4, 2023.
- davanstrien. davanstrien/dataset-tldr-preference-dpo. <https://huggingface.co/datasets/davanstrien/dataset-tldr-preference-dpo>, 2024.
- Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation, 2025. URL <https://arxiv.org/abs/2506.01937>.

- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. Vl-rewardbench: A challenging benchmark for vision-language generative reward models, 2025b. URL <https://arxiv.org/abs/2411.17451>.
- Yongchao Chen, Yilun Hao, Yueying Liu, Yang Zhang, and Chuchu Fan. Codesteer: Symbolic-augmented language models via code/text guidance, 2025. URL <https://arxiv.org/abs/2502.04350>.
- Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Mask-dpo: Generalizable fine-grained factuality alignment of llms, 2025. URL <https://arxiv.org/abs/2503.02846>.
- Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation, 2025b. URL <https://arxiv.org/abs/2503.05236>.
- Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following, 2025. URL <https://arxiv.org/abs/2504.07957>.
- Tianduo Wang, Shichen Li, and Wei Lu. Self-training with direct preference optimization improves chain-of-thought reasoning, 2024c. URL <https://arxiv.org/abs/2407.18248>.
- Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.539. URL <https://aclanthology.org/2024.acl-long.539>.
- Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. Self-play with execution feedback: Improving instruction-following capabilities of large language models, 2024. URL <https://arxiv.org/abs/2406.13542>.
- Jiuding Yang, Weidong Guo, Kaitong Yang, Xiangyang Li, Zhiwei Rao, Yu Xu, and Di Niu. Enhancing and assessing instruction-following with fine-grained instruction variants, 2024b. URL <https://arxiv.org/abs/2406.11301>.
- Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. Conifer: Improving complex constrained instruction-following ability of large language models, 2024. URL <https://arxiv.org/abs/2404.02823>.
- Junru Lu, Siyu An, Min Zhang, Yulan He, Di Yin, and Xing Sun. Fipo: Free-form instruction-oriented prompt optimization with preference dataset and modular fine-tuning schema, 2024c. URL <https://arxiv.org/abs/2402.11811>.
- Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. Rs-dpo: A hybrid rejection sampling and direct preference optimization method for alignment of large language models, 2024. URL <https://arxiv.org/abs/2402.10038>.
- Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen. Flame: Factuality-aware alignment for large language models, 2024c. URL <https://arxiv.org/abs/2405.01525>.
- Binjie Wang, Steffi Chern, Ethan Chern, and Pengfei Liu. Halu-j: Critique-based hallucination judge, 2024d. URL <https://arxiv.org/abs/2407.12943>.
- Rachneet Sachdeva, Yixiao Song, Mohit Iyyer, and Iryna Gurevych. Fine-grained hallucination detection and mitigation in long-form question answering, 2024. URL <https://arxiv.org/abs/2407.11930>.
- Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality, 2023. URL <https://arxiv.org/abs/2311.08401>.
- Leonidas Gee, Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. Code-optimise: Self-generated preference data for correctness and efficiency, 2024. URL <https://arxiv.org/abs/2406.12502>.
- Nikhil Pinnaparaju, Reshith Adithyan, Duy Phung, Jonathan Tow, James Baicoianu, Ashish Datta, Maksym Zhuravinsky, Dakota Mahan, Marco Bellagente, Carlos Riquelme, and Nathan Cooper. Stable code technical report, 2024. URL <https://arxiv.org/abs/2404.01226>.
- Daniel Nichols, Pranav Polasam, Harshitha Menon, Aniruddha Marathe, Todd Gamblin, and Abhinav Bhatele. Performance-aligned llms for generating fast code, 2024. URL <https://arxiv.org/abs/2404.18864>.
- Zhaofeng Liu, Jing Su, Jia Cai, Jingzhi Yang, and Chenfan Wu. Instruct-code-llama: Improving capabilities of language model in competition level code generation by online judge feedback. In De-Shuang Huang, Zhanjun Si, and Qinhu Zhang, editors, *Advanced Intelligent Computing Technology and Applications*, pages 127–137, Singapore, 2024e. Springer Nature Singapore. ISBN 978-981-97-5669-8.

- Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning, 2024b.
- Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward, 2024b. URL <https://arxiv.org/abs/2404.01258>.
- Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios, 2024. URL <https://arxiv.org/abs/2403.04640>.
- 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. URL <https://arxiv.org/abs/2311.12908>.
- Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model, 2024c. URL <https://arxiv.org/abs/2311.13231>.
- Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization, 2024. URL <https://arxiv.org/abs/2404.09956>.
- Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F. Chen. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization, 2024. URL <https://arxiv.org/abs/2409.10157>.
- Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechalign: Aligning speech generation to human preferences, 2024c. URL <https://arxiv.org/abs/2404.05600>.
- Hyundong Cho, Nicolaas Jedema, Leonardo F. R. Ribeiro, Karishma Sharma, Pedro Szekely, Alessandro Moschitti, Ruben Janssen, and Jonathan May. Speechworthy instruction-tuned language models, 2024. URL <https://arxiv.org/abs/2409.14672>.
- Jenny Sheng, Matthieu Lin, Andrew Zhao, Kevin Pruvost, Yu-Hui Wen, Yangguang Li, Gao Huang, and Yong-Jin Liu. Exploring text-to-motion generation with human preference, 2024. URL <https://arxiv.org/abs/2404.09445>.
- Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. On softmax direct preference optimization for recommendation, 2024c. URL <https://arxiv.org/abs/2406.09215>.
- Zhuoxi Bai, Ning Wu, Fengyu Cai, Xinyi Zhu, and Yun Xiong. Finetuning large language model for personalized ranking, 2024. URL <https://arxiv.org/abs/2405.16127>.
- Junjie Hu, Peng Wu, Shiyi Wang, Binju Wang, and Guang Yang. A human feedback strategy for photoresponsive molecules in drug delivery: Utilizing gpt-2 and time-dependent density functional theory calculations. *Pharmaceutics*, 16(8), 2024. ISSN 1999-4923. doi: 10.3390/pharmaceutics16081014. URL <https://www.mdpi.com/1999-4923/16/8/1014>.
- Hong Nguyen, Hoang Nguyen, Melinda Chang, Hieu Pham, Shrikanth Narayanan, and Michael Pazzani. Conpro: Learning severity representation for medical images using contrastive learning and preference optimization, 2024. URL <https://arxiv.org/abs/2404.18831>.
- Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via direct preference optimization. *bioRxiv*, 2024. doi: 10.1101/2024.05.20.595026. URL <https://www.biorxiv.org/content/early/2024/05/21/2024.05.20.595026>.
- 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, 1952b. ISSN 00063444, 14643510. URL <http://www.jstor.org/stable/2334029>.
- Zicheng Lin, Tian Liang, Jiahao Xu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhance llm's reasoning capability. *arXiv preprint arXiv:2411.19943*, 2024d.
- Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Hassan Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q^* -approximation for sample-efficient RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=QYigQ6gXNw>.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei.

- Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/ARXIV.2210.11416. URL <https://doi.org/10.48550/arXiv.2210.11416>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.
- William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022. URL <http://jmlr.org/papers/v23/21-0998.html>.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *CoRR*, abs/2303.18223, 2023. doi: 10.48550/ARXIV.2303.18223. URL <https://doi.org/10.48550/arXiv.2303.18223>.
- Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=tPNH0oZF19>.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022b. doi: 10.48550/ARXIV.2204.05862. URL <https://doi.org/10.48550/arXiv.2204.05862>.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif: Scaling reinforcement learning from human feedback with ai feedback, 2023.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017b.
- 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/mnih16.html>.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL <https://arxiv.org/abs/1801.01290>.
- Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL <http://arxiv.org/abs/1909.08593>.
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020. URL <https://arxiv.org/abs/2009.01325>.
- Eric J. Michaud, Adam Gleave, and Stuart Russell. Understanding learned reward functions. *CoRR*, abs/2012.05862, 2020. URL <https://arxiv.org/abs/2012.05862>.
- Bo Wang, Qinyuan Cheng, Runyu Peng, Rong Bao, Peiji Li, Qipeng Guo, Linyang Li, Zhiyuan Zeng, Yunhua Zhou, and Xipeng Qiu. Implicit reward as the bridge: A unified view of sft and dpo connections, 2025c. URL <https://arxiv.org/abs/2507.00018>.
- Yuzi Yan, Yibo Miao, Jialian Li, Yipin Zhang, Jian Xie, Zhijie Deng, and Dong Yan. 3d-properties: Identifying challenges in DPO and charting a path forward. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=9Hxdixed7p>.

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992. URL <https://api.semanticscholar.org/CorpusID:8456150>.
- Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment, 2024b. URL <https://arxiv.org/abs/2405.00675>.
- Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences, 2024. URL <https://arxiv.org/abs/2404.03715>.
- Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback, 2024. URL <https://arxiv.org/abs/2312.00886>.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024.
- Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^* : Your language model is secretly a q-function. *CoRR*, abs/2404.12358, 2024b. doi: 10.48550/ARXIV.2404.12358. URL <https://doi.org/10.48550/arXiv.2404.12358>.
- Joey Hejna, Rafael Raffailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. Contrastive preference learning: Learning from human feedback without reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=iX1RjVQODj>.
- Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training, 2019. URL <https://arxiv.org/abs/1908.04319>.
- Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024e.
- Leonard Adolphs, Tianyu Gao, Jing Xu, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. The cringe loss: Learning what language not to model, 2022. URL <https://arxiv.org/abs/2211.05826>.
- 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.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *Forty-first International Conference on Machine Learning*, 2024d.
- Haoqin Tu, Bingchen Zhao, Chen Wei, and Cihang Xie. Sight beyond text: Multi-modal training enhances llms in truthfulness and ethics, 2023. URL <https://arxiv.org/abs/2309.07120>.
- AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1, 2024.
- Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024c.
- Jondurbin. Jondurbin/gutenberg-dpo-v0.1. <https://huggingface.co/jondurbin/bagel-7b-v0.1>, 2024b.
- Yuasosnin. Yuasosnin/imdb-dpo. <https://huggingface.co/jondurbin/bagel-7b-v0.1>, 2024.
- Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. Huggingface h4 stack exchange preference dataset, 2023. URL <https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences>.
- Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL <https://aclanthology.org/W17-4508>.
- RyokoAI. Ryokoai/sharegpt52k. <https://huggingface.co/datasets/RyokoAI/ShareGPT52K>, 2023.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL <https://arxiv.org/abs/2209.07858>.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL <https://arxiv.org/abs/1405.0312>.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023a. URL <https://arxiv.org/abs/2311.12793>.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding, 2020. URL <https://arxiv.org/abs/2007.10937>.

Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2572–2581, 2020. doi: 10.1109/CVPR42600.2020.00265.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL <https://arxiv.org/abs/1612.00837>.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. URL <https://arxiv.org/abs/1906.00067>.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. Chatgpt: Applications, opportunities, and threats, 2023. URL <https://arxiv.org/abs/2304.09103>.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Biket, 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, 2023b.

AI@Meta. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

"Teknium""theemozilla" "karan4d" "huemin_art". Nous hermes 2 mixtral 8x7b dpo, 2024. URL [<https://huggingface.co/NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO>] (<https://huggingface.co/NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO>).

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tat-sunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023b.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xtest: A test suite for identifying exaggerated safety behaviours in large language models, 2024. URL <https://arxiv.org/abs/2308.01263>.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms, 2023d. URL <https://arxiv.org/abs/2308.13387>.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL <https://arxiv.org/abs/2305.20050>.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. Octopack: Instruction tuning code large language models, 2024. URL <https://arxiv.org/abs/2308.07124>.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL <https://arxiv.org/abs/2210.08402>.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricu, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, and Balaji. Gemini: A family of highly capable multimodal models, 2024. URL <https://arxiv.org/abs/2312.11805>.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023b.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023. URL <https://arxiv.org/abs/2301.13688>.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions, 2023. URL <https://arxiv.org/abs/2304.12244>.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024d. URL <https://arxiv.org/abs/2407.10671>.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL <https://arxiv.org/abs/2311.07911>.

DeepSeek-AI, ;, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL <https://arxiv.org/abs/2401.02954>.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024a. URL <https://arxiv.org/abs/2401.04088>.

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*, 2023c.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze

- Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. Baichuan 2: Open large-scale language models, 2023. URL <https://arxiv.org/abs/2309.10305>.
01. AI, :, Alex Young, Bei Chen, Chao Li, Chengan Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024. URL <https://arxiv.org/abs/2403.04652>.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations, 2023b. URL <https://arxiv.org/abs/2310.20246>.
- Renze Lou, Kai Zhang, and Wenpeng Yin. Large language model instruction following: A survey of progresses and challenges, 2024b. URL <https://arxiv.org/abs/2303.10475>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022b. URL <https://arxiv.org/abs/2203.02155>.
- Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations?, 2024. URL <https://arxiv.org/abs/2405.05904>.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
- Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, and Fei Wu. LoraRetriever: Input-aware LoRA retrieval and composition for mixed tasks in the wild. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 4447–4462, Bangkok, Thailand and virtual meeting, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024. findings-acl.263. URL <https://aclanthology.org/2024.findings-acl.263>.
- Sewon Min, Kalpesh Krishna, Xinxin Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023. URL <https://arxiv.org/abs/2305.14251>.
- Peifeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. Direct judgement preference optimization, 2024e. URL <https://arxiv.org/abs/2409.14664>.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024f.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023b.
- Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024d. URL <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024. URL <https://arxiv.org/abs/2407.10759>.
- Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference, 2024e. URL <https://arxiv.org/abs/2402.08265>.
- Daria Diatlova and Vitaly Shutov. Emospeech: Guiding fastspeech2 towards emotional text to speech. *arXiv preprint arXiv:2307.00024*, 2023.
- Xiang Li, Zhi-Qi Cheng, Jun-Yan He, Xiaojiang Peng, and Alexander G Hauptmann. Mm-tts: A unified framework for multimodal, prompt-induced emotional text-to-speech synthesis. *arXiv preprint arXiv:2404.18398*, 2024b.
- Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024b.

- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision, 2023. URL <https://arxiv.org/abs/2312.09390>.
- Shuai Zhao, Leilei Gan, Zhongliang Guo, Xiaobao Wu, Luwei Xiao, Xiaoyu Xu, Cong-Duy Nguyen, and Luu Anh Tuan. Weak-to-strong backdoor attacks for llms with contrastive knowledge distillation. *arXiv preprint arXiv:2409.17946*, 2024b.
- Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024.
- Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024d.
- Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint, 2024. URL <https://arxiv.org/abs/2312.11456>.
- Lei Li, Yekun Chai, Shuhuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. Tool-augmented reward modeling. In *The Twelfth International Conference on Learning Representations*, 2024c. URL <https://openreview.net/forum?id=d94x0gWTUX>.
- Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=E01k9048soZ>.
- Wanggui He, Siming Fu, Mushui Liu, Xierui Wang, Wenyi Xiao, Fangxun Shu, Yi Wang, Lei Zhang, Zhelun Yu, Haoyuan Li, et al. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. *arXiv preprint arXiv:2407.07614*, 2024.
- Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. A survey on long video generation: Challenges, methods, and prospects, 2024d. URL <https://arxiv.org/abs/2403.16407>.
- Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.