

# A Survey of Direct Preference Optimization

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, Yongbin Li, and Dacheng Tao, *Fellow, IEEE*

**Abstract**—Large Language Models (LLMs) have demonstrated unprecedented generative capabilities, yet their alignment with human values remains critical for ensuring helpful and harmless deployments. While Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful paradigm for aligning LLMs with human preferences, its reliance on complex reward modeling introduces inherent trade-offs in computational efficiency and training stability. In this context, Direct Preference Optimization (DPO) has recently gained prominence as a streamlined alternative that directly optimizes LLMs using human preferences, thereby circumventing the need for explicit reward modeling. Owing to its theoretical elegance and computational efficiency, DPO has rapidly attracted substantial research efforts exploring its various implementations and applications. However, this field currently lacks systematic organization and comparative analysis. In this survey, we conduct a comprehensive overview of DPO and introduce a novel taxonomy, categorizing previous works into four key dimensions: *data strategy*, *learning framework*, *constraint mechanism*, and *model property*. We further present a rigorous empirical analysis of DPO variants across standardized benchmarks. Additionally, we discuss real-world applications, open challenges, and future directions for DPO. This work delivers both a conceptual framework for understanding DPO and practical guidance for practitioners, aiming to advance robust and generalizable alignment paradigms. All collected resources are available and will be continuously updated at <https://github.com/liushunyu/awesome-direct-preference-optimization>.

**Index Terms**—Alignment, Direct Preference Optimization, Large Language Models, Reinforcement Learning from Human Feedback.

## 1 INTRODUCTION

THE rapid advancement of Large Language Models (LLMs) has revolutionized artificial intelligence [1, 2, 3, 4, 5, 6, 7, 8], enabling unprecedented generative capabilities across diverse applications, such as dialogue systems [9, 10], code generation [11, 12, 13], and medical diagnosis [14, 15, 16, 17]. Models like OpenAI-o1 [18] and DeepSeek-R1 [19] have demonstrated remarkable proficiency in understanding and generating human-like text, outperforming traditional language processing techniques [20]. However, their immense power also introduces significant risks: LLMs may inadvertently produce harmful content (*e.g.*, jailbreak suggestion) [21], exhibit hallucination behaviors (*e.g.*, misinformation) [22], or propagate sociocultural stereotypes (*e.g.*, biased recommendations) [23]. Ensuring that these models align with human values (producing outputs that are helpful, harmless, and honest) has thus become a cornerstone of responsible AI development [24].

The critical challenge of aligning LLMs with human values stems from the inherent complexity of encoding abstract

ethical principles into concrete model behaviors [25, 26, 27]. Traditional approaches, such as rule-based filtering or supervised learning on curated datasets, often prove inadequate due to their inability to generalize across diverse contexts and adapt to evolving societal norms [28]. The emergence of preference-based alignment paradigms addresses these limitations by framing the problem as optimizing for human feedback rather than inflexible heuristics [29, 30, 31, 32]. This shift recognizes that LLM decision-making often involves nuanced trade-offs between competing values, requiring flexible frameworks capable of incorporating subjective human preferences [33].

Building upon these insights, Reinforcement Learning from Human Feedback (RLHF) [34, 35] has emerged as the predominant alignment paradigm, leveraging human preferences to guide model optimization. In the RLHF pipeline, human annotators first rank the outputs generated by the language model, and these comparisons are used to train a reward model that quantifies human preferences. The language model is then fine-tuned using RL guided by this reward model, enabling the language model to align with human values by maximizing the predicted rewards. The success of RLHF in aligning models like ChatGPT [36, 37] and Claude [38, 39] underscores its practical utility. By translating subjective human preferences into an objective reward signal, RLHF facilitates the optimization of model behavior for value alignment. However, this RLHF paradigm suffers from critical limitations of computational complexity and training instability. Training a separate reward model demands substantial computational resources and high-quality human preference data, which scales poorly across different domains. Moreover, the RL phase often struggles with optimization challenges, such as reward hacking [40] and mode collapse [41].

These limitations have spurred interest in alternative

*This research is supported by the RIE2025 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A\*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). (Corresponding author: Dacheng Tao.)*

*Shunyu Liu, Junjie Zhang, Rongcheng Tu and Dacheng Tao are with Nanyang Technological University, Singapore (e-mail: shunyu.liu@ntu.edu.sg; junjie.zhang@ntu.edu.sg; turongcheng@gmail.com; dacheng.tao@ntu.edu.sg). Wenkai Fang, Yang Zhou, Kongcheng Zhang, and Mingli Song are with the College of Computer Science and Technology, Zhejiang University, China (e-mail: wenkai.fang@zju.edu.cn; imzhouyang@zju.edu.cn; zhangkc@zju.edu.cn; brooksong@zju.edu.cn).*

*Zetian Hu is with the School of Aerospace Engineering, Tsinghua University, China (e-mail: huzt22@mails.tsinghua.edu.cn).*

*Ting-En Lin, Fei Huang, and Yongbin Li are with the Tongyi Lab, Alibaba Group, China (e-mail: ting-en.lte@alibaba-inc.com; f.huang@alibaba-inc.com; shuide.lyb@alibaba-inc.com).*

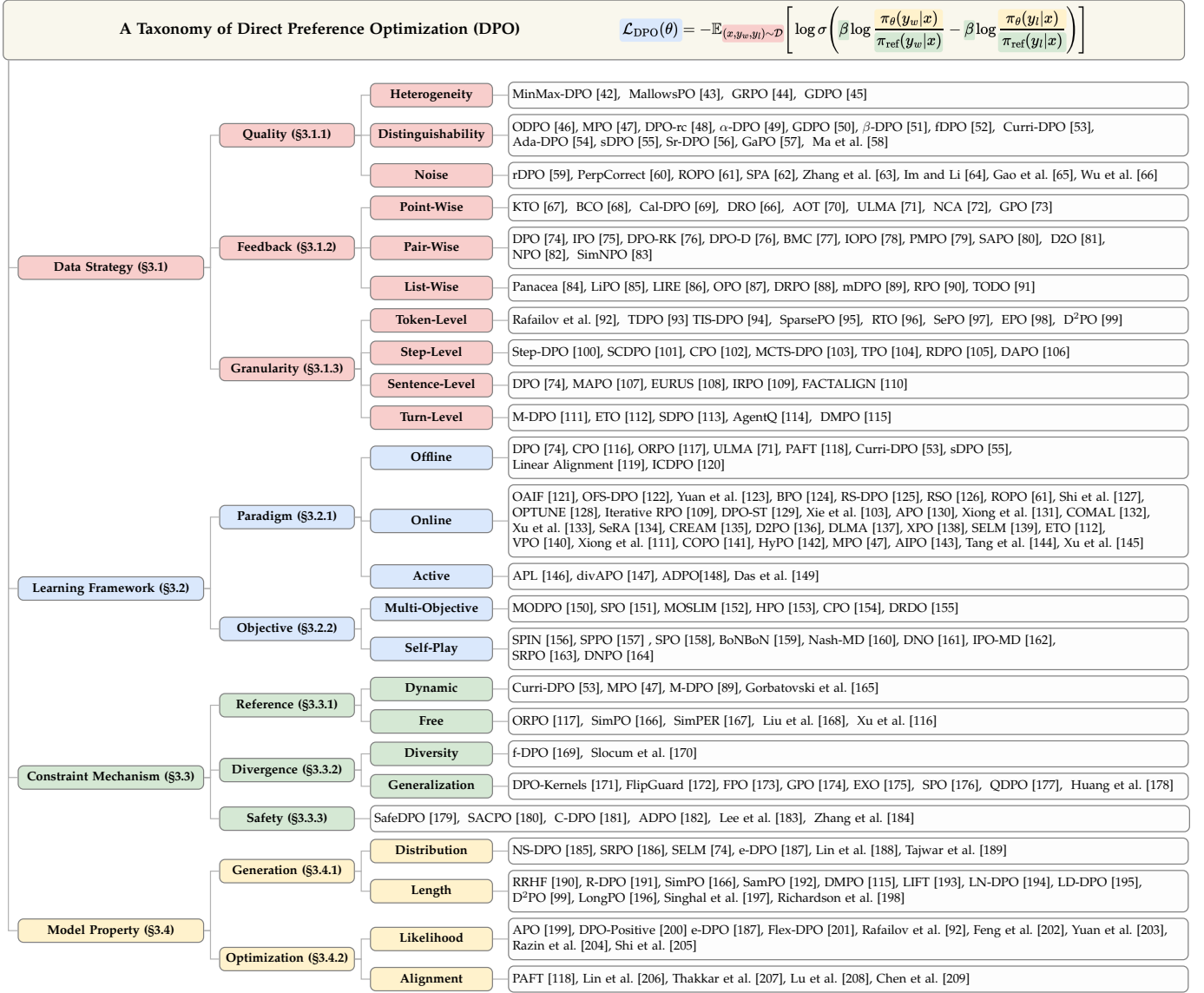


Fig. 1: A taxonomy of DPO. We categorize existing DPO works into four branches: *data strategy*, *learning framework*, *constraint mechanism*, and *model property*. Different colored boxes indicate different categories and their corresponding representative references.

alignment methods that bypass reward modeling while preserving the benefits of preference-based learning. Direct Preference Optimization (DPO) [74, 210] represents a groundbreaking shift in this direction. Unlike RLHF, DPO re-frames alignment as a supervised learning problem, directly optimizing the LLM policy using preference data without explicit reward modeling. By leveraging a closed-form mapping between reward functions and optimal policies, DPO eliminates the need for iterative RL training, reducing computational overhead and improving stability. Due to its inherent advantages, DPO has rapidly gained increasing attention from research communities. Existing studies vary widely in data strategies (*e.g.*, point-wise *v.s.* pair-wise feedback) [67, 211], learning frameworks (*e.g.*, offline *v.s.* online learning) [121, 122, 126], constraint mechanisms (*e.g.*, different divergence constraints) [169, 171], and model properties (*e.g.*, length bias) [191, 195]. Recent advancements in DPO variants have demonstrated remarkable efficacy in enhancing model alignment with human preferences, achieving unprecedented success across diverse domains [32].

These developments position DPO-based approaches as a compelling alternative to conventional RLHF paradigms for preference alignment tasks. However, despite its promise, the DPO research landscape remains fragmented.

Several surveys related to DPO have been published in recent years, yet they exhibit notable limitations in their scope and analysis of DPO. (1) *Scope limitations*. While an early survey of [212] presents a comprehensive overview of preference-based RL methods, it predates the advent of DPO and does not address its applications to modern LLMs. Recent surveys on alignment [24, 26, 213, 214] provide broad overviews of LLM alignment techniques but only offer cursory summaries of DPO-related approaches without in-depth analysis. Similarly, surveys on learning from human feedback [30, 215, 216, 217] also only briefly mention DPO as a potential alternative. (2) *Taxonomy deficiencies*. Gao et al. [29] and Winata et al. [32] introduce a simplified taxonomy for preference learning, while overlooking technical distinctions within its broad categorization. In contrast, Wang et al. [31] attempt to classify preference learning across dimensions

such as reinforcement learning, reward modeling, feedback, and optimization. However, this taxonomy suffers from significant conceptual overlaps (e.g. reinforcement learning inherently involves optimization). A recent work by Xiao et al. [210] categorizes DPO studies through isolated research questions, which, while useful for problem identification, fragments the methodological connections. Our survey addresses these gaps by presenting the first comprehensive analysis specifically focused on DPO. The main contributions of this survey are summarized as follows:

- In this survey, we introduce a novel taxonomy that categorizes existing DPO works into four key dimensions based on different components of the DPO loss: *data strategy*, *learning framework*, *constraint mechanism*, and *model property*, as shown in Fig. 1. This taxonomy provides a systematic framework for understanding the methodological evolution of DPO and highlights the key distinctions between different variations.
- We conduct a rigorous empirical analysis of DPO variants across standardized benchmarks, revealing critical insights into their performance in diverse scenarios. This analysis offers a comprehensive evaluation of DPO variants and provides practical guidance for practitioners.
- We discuss real-world applications of DPO and highlight its potential to democratize alignment research by enabling efficient and scalable preference learning across diverse domains. We also outline open challenges and future directions for DPO research, emphasizing the need for robust and generalizable alignment paradigms.

The remainder of this survey is organized as follows. Section 2 introduces the background and formulation of DPO. Section 3 presents a taxonomy of DPO, categorizing existing works based on key dimensions. Section 4 describes standardized benchmarks for evaluating DPO methods and presents empirical results. Section 5 discusses real-world applications of DPO and highlights its potential. Section 6 outlines open challenges and future directions for DPO research. Finally, Section 7 concludes the survey.

## 2 BACKGROUND AND FORMULATION

Preference learning aims to train language model policies to generate responses that better align with human preferences. Specifically, we denote the language model policy as  $\pi(y|x)$ , where  $x$  represents the input prompt and  $y$  is a candidate response (completion). A language model can be viewed as an autoregressive function that sequentially predicts tokens based on prior context. Mathematically, this is expressed as:  $\pi(y|x) = \prod_{t=1}^T \pi(y_t|y_{<t}, x)$ . where  $y = (y_1, y_2, \dots, y_T)$  is the response sequence,  $y_t$  represents the token at position  $t$ ,  $y_{<t} = (y_1, y_2, \dots, y_{t-1})$  denotes the sequence of previously generated tokens, and  $\pi(y_t|y_{<t}, x)$  is the probability of generating token  $y_t$  conditioned on both the input  $x$  and the previously generated tokens  $y_{<t}$ . In the context of preference learning, the preference data is defined as a collection of triplets:  $\mathcal{D} = \{(x, y_w, y_l)\}$ , where  $x$  is an input prompt, and  $y_w$  and  $y_l$  are two candidate responses, with  $y_w$  being preferred over  $y_l$  (denoted as  $y_w \succ y_l$ ). The responses  $y_w$  and  $y_l$  are commonly referred to as the chosen (winning) and rejected (losing) responses, respectively.

To leverage preference data  $\mathcal{D}$  for training the language model policy  $\pi$ , RLHF employs a two-stage process that first learns a reward function from preference data and then optimizes the policy using RL [34, 35, 36]. In contrast, DPO directly optimizes the policy using preference data, eliminating the need for an explicit reward model [74]. The following sections provide a detailed formulation of RLHF and DPO, highlighting their key differences and advantages. Moreover, we also introduce several preference optimization methods that are concurrent with DPO.

### 2.1 Reinforcement Learning from Human Feedback

RLHF formulates preference learning as a two-stage process that involves reward modeling and policy optimization. Typically, the RLHF process of LLMs also includes Supervised Fine-Tuning (SFT) prior to these stages, where high-quality demonstration data is used to fine-tune the pre-trained language model to obtain the SFT model  $\pi_{\text{sft}}$ , establishing instruction-following capabilities to support subsequent preference learning [35, 36].

#### 2.1.1 Reward Modeling

In the reward modeling stage, the goal is to learn a separate reward model  $r_\phi$  parameterized by  $\phi$ , which quantifies how well a response  $y$  satisfies human preference for a given prompt  $x$ . Using the Bradley-Terry model [218], the preference probability that response  $y_w$  is preferred over response  $y_l$  for prompt  $x$  is modeled as follows:

$$P(y_w \succ y_l | x) = \frac{\exp(r_\phi(x, y_w))}{\exp(r_\phi(x, y_w)) + \exp(r_\phi(x, y_l))}. \quad (1)$$

The reward model is trained by minimizing the negative log-likelihood of Eq. 1 as the loss function:

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

where  $\sigma(\cdot)$  denotes the sigmoid function. This objective encourages the model to assign higher rewards to responses that are preferred by humans.

#### 2.1.2 Policy Optimization

After training the reward model, the next stage is to optimize the language model policy  $\pi_\theta$  parameterized by  $\theta$  using RL. This policy  $\pi_\theta$  is initialized by the SFT model  $\pi_{\text{sft}}$ . We use the learned reward model  $r_\phi$  to provide feedback that guides the policy  $\pi_\theta$  to generate responses with higher rewards. The optimization objective is defined as follows:

$$J_\pi(\theta) = \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \left[ r_\phi(x, y) - \beta \log \frac{\pi_\theta(\cdot|x)}{\pi_{\text{ref}}(\cdot|x)} \right], \quad (3)$$

where  $\beta > 0$  is a hyperparameter that controls the strength of the Kullback-Leibler (KL) divergence penalty. Here, the term  $\log \pi_\theta(\cdot|x) / \pi_{\text{ref}}(\cdot|x)$  represents the KL divergence between the current policy  $\pi_\theta$  and a reference policy  $\pi_{\text{ref}}$ . In practice, the reference policy  $\pi_{\text{ref}}$  is set to the SFT model  $\pi_{\text{sft}}$ , ensuring that the updated policy remains close to the initial model.

To optimize the above objective, Proximal Policy Optimization (PPO) [219] has emerged as a promising RL algorithm for LLMs. PPO stabilizes training by constraining policy updates within a trust region via a clipped objective, which prevents significant deviations from the previous

policy. However, PPO requires an additional critic model to estimate value functions for advantage calculation, thereby introducing extra computational and memory overhead. To address this, recent methods, such as RLOO [220], ReMax [221], GRPO [222], and Reinforce++ [223], introduce critic-free advantage estimation to reduce resource demands while maintaining stable optimization, making them more scalable for large-scale LLM training.

## 2.2 Direct Preference Optimization

DPO offers an alternative that streamlines the training process by directly optimizing the policy with preference data [74, 224, 225, 226, 227, 228, 229], thereby eliminating the need for explicit reward modeling in RLHF. The key idea of DPO is a closed-form solution of Eq. 3 that connects reward with the optimal policies. Specifically, the optimal policy corresponding to a given  $r$  is defined as follows:

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

where the partition function  $Z(x)$  is defined as:

$$Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right). \quad (5)$$

By rearranging the above equation, the reward  $r$  can be recovered from the optimal policy  $\pi^*$ :

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

Notice that the partition function  $Z(x)$  depends only on the prompt  $x$ . By substituting this expression into the preference model of Eq. 1, the preference probability model that  $y_w$  is preferred over  $y_l$  becomes:

$$P(y_w \succ y_l|x) = \sigma\left(\beta \log \frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right). \quad (7)$$

Based on the above preference probability model, DPO directly optimizes the language model policy  $\pi_\theta$  by minimizing the following negative log-likelihood loss function:

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

where the KL constraint is implicitly integrated through the use of the reference model  $\pi_{\text{ref}}$ . By minimizing this DPO loss, we directly train the policy to satisfy human preferences without resorting to a separate reward modeling stage or using reinforcement learning optimization as in RLHF, significantly reducing implementation complexity while improving training stability.

## 2.3 Other Preference Optimization

In addition to DPO, several concurrent preference optimization methods [190, 230, 231] have been proposed that offer alternative approaches to RLHF. These methods explore different strategies for optimizing LLMs to align with human preference without RL. Below, we provide a brief introduction to these approaches.

### 2.3.1 Sequence Likelihood Calibration

Zhao et al. [230] propose Sequence Likelihood Calibration with Human Feedback (SLiC-HF) to directly align LLMs with human preferences. Specifically, the loss function of SLiC-HF is defined as follows:

$$\mathcal{L}_{\text{SLiC-HF}}(\theta) = \max(0, \delta - \log \pi_\theta(y_w|x) + \log \pi_\theta(y_l|x)) - \lambda \log \pi_\theta(y^*|x), \quad (9)$$

where the first term is the rank calibration loss with  $\delta$  as a margin hyperparameter, and the second term is the cross-entropy regularization loss with  $\lambda$  as a regularization weight.  $y^*$  is obtained from either high-quality supervised responses in the SFT dataset or the top-ranked candidate response generated by the SFT model.

### 2.3.2 Rank Responses to Align Human Feedback

Yuan et al. [190] introduce Rank Responses to align Human Feedback (RRHF) for LLMs. RRHF extends pair-wise ranking by considering the list-wise ranking order of multiple responses, thus better utilizing the preference information. For an input prompt  $x$  and  $N$  candidate responses  $\{y_i\}_{i=1}^N$ , it optimizes the model to assign higher probabilities to higher-ranked responses via a ranking loss and directly supervises the best response using cross-entropy as follows:

$$\mathcal{L}_{\text{RRHF}}(\theta) = \sum_{r_i < r_j} \max\left(0, \frac{\log \pi_\theta(y_i|x)}{\|y_i\|} - \frac{\log \pi_\theta(y_j|x)}{\|y_j\|}\right) - \lambda \log \pi_\theta(y^*|x), \quad (10)$$

where  $r_i = r_\phi(x, y_i)$  represents the reward of the response  $y_i$  and  $y^* = \arg \max_{y_i} r_i$  is the response with the highest reward. Although RRHF avoids the need for reinforcement learning in RLHF, it still utilizes a reward model  $r_\phi$  to rank candidate responses based on human preferences.

### 2.3.3 Preference Ranking Optimization

Similarly, Song et al. [231] propose Preference Ranking Optimization (PRO) to align LLMs with human preferences by leveraging multiple responses  $\{y_i\}_{i=1}^N$  with the human-annotated order  $y_1 \succ y_2 \succ \dots \succ y_N$ . The loss function of PRO is defined as follows:

$$\mathcal{L}_{\text{PRO}}(\theta) = - \sum_{i=1}^{N-1} \log \frac{\exp\left(\frac{1}{\|y_i\|} \log \pi_\theta(y_i|x) / \mathcal{T}_i^i\right)}{\sum_{j=i}^N \exp\left(\frac{1}{\|y_j\|} \log \pi_\theta(y_j|x) / \mathcal{T}_i^j\right)}, \quad (11)$$

where the dynamic penalty temperature is defined as  $\mathcal{T}_i^j = 1/(r_\phi(x, y^j) - r_\phi(x, y^i))$  and  $\mathcal{T}_i^i = \min_{i < j} \mathcal{T}_i^j$ . This temperature ensures that the probability gap between higher-ranked and lower-ranked responses is adaptively scaled according to their reward differences, thereby stabilizing the optimization process.

## 3 A TAXONOMY OF DPO

In this section, we introduce a novel taxonomy that categorizes existing DPO works based on four key dimensions: *data strategy*, *learning framework*, *constraint mechanism*, and *model property*. As illustrated in Fig. 1, these four dimensions are derived from different components of the DPO loss, providing a systematic framework for understanding the methodological evolution of DPO and highlighting the key distinctions between different variations.

### 3.1 Data Strategy of DPO

The data strategy constitutes the foundational pillar of DPO, focusing on how to leverage diverse types of preference data for training LLMs. As shown in Fig. 2, our taxonomy identifies three principal axes of data strategy: quality, feedback, and granularity.

#### 3.1.1 Data Quality

The quality of preference data is a critical factor in determining the effectiveness of DPO training. High-quality data ensures that LLMs effectively learn to align with human preferences, while low-quality data may introduce noise and bias, leading to suboptimal model performance. We categorize data quality considerations into three key aspects: heterogeneity, distinguishability, and noise.

**(a) Data Heterogeneity.** Conventional DPO methods assume uniform human preferences when annotating data, thereby overlooking the diversity among annotators. This assumption often skews the model toward the preferences of the majority while neglecting minority viewpoints, potentially leading to biases and unfair treatment of underrepresented groups. To address this issue, Chidambaram et al. [42] propose EM-DPO, which learns the distribution of different preference types and their corresponding response strategies. Building on this, they introduce the MinMax-DPO algorithm, which selects a strategy by minimizing the maximum regret across subgroups, ensuring a more balanced representation of preferences among all groups. MallowsPO [43] decomposes the implicit rewards in DPO into prompt dispersion and response scaling rewards. It introduces a novel objective function to capture human preferences for diverse responses to the same prompt. GRPO [44] formulates an objective function that minimizes the loss for the worst-case group, thereby ensuring fairness by prioritizing the disadvantaged groups in the optimization process. GDPO [45] models the language generation process as a combination of belief distribution prediction and belief-based response generation. The corresponding GDPO loss function consists of belief calibration loss and belief-conditioned preference alignment loss. The former encourages the model to capture the diversity of beliefs across groups, while the latter ensures that generated responses align with the given belief.

**(b) Data Distinguishability.** A key limitation of DPO is its inability to account for the distinguishability of preference between responses [46, 50, 51, 56, 57]. In some cases, the preferred response is only marginally better than the dispreferred one, while in others, the dispreferred response contains harmful or misleading content, making it significantly worse. Thus, optimization should focus more on cases with substantial preference differences while reducing the effort spent on minor differences. However, most existing methods treat all samples equally, ignoring this data distinguishability. To address this, ODPO [46] introduces a monotonically increasing offset function, requiring the reward of the preferred response to exceed that of the dispreferred one by a certain margin. This ensures stronger updates for larger preference gaps. Similarly, Ada-DPO [54] introduces an instance-specific nonlinear scaling parameter, assigning larger weights to strong preference pairs and smaller weights to ambiguous ones based on the reward differences, thereby capturing

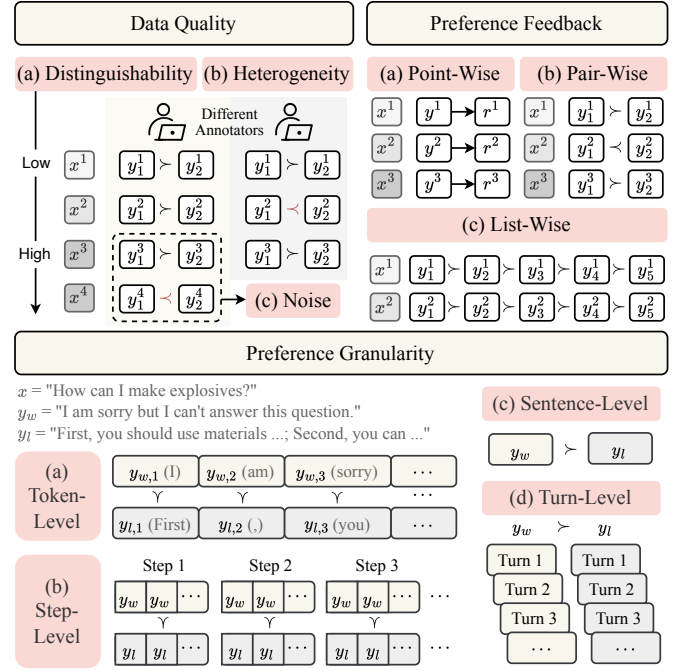


Fig. 2: An overview of DPO data strategy.

different levels of data distinguishability. DPO-rc [48] also incorporates the preference reward difference as a coefficient in the loss function.  $\alpha$ -DPO [49] introduces an adaptive preference distribution to obtain dynamic reward margins based on the distribution difference between the policy and reference models.  $\beta$ -DPO [51] analyzes the optimal  $\beta$  parameter for datasets with different reward margins, which dynamically adjusts  $\beta$  based on batch-level reward differences. They also introduce  $\beta$ -guided data filtering to prioritize valuable training data. Curri-DPO [53] sorts preference pairs by reward differences and trains progressively from large to small differences, enabling curricular learning. Similarly, MPO [47] utilizes a reward model to score responses generated by the SFT model, constructing a preference dataset and partitioning it based on preference differences to learn from simple to complex tasks. sDPO [55] computes reward accuracy for different datasets based on an initial target model and partitions the dataset in descending order of accuracy, allowing the model to first optimize on simpler samples. Ma et al. [58] propose a preference dataset construction method that adjusts update weights based on response accuracy, assigning lower weights when the model demonstrates higher proficiency. Furthermore, fDPO [52] enhances DPO training by filtering out samples where the generated response of the model policy surpasses the preferred dataset response in reward score.

**(c) Data Noise.** Human-generated preference annotations often contain inconsistencies, errors, or noise, negatively affecting the performance of DPO. Such noisy data can mislead models, impairing their ability to accurately capture true preferences and generalize effectively to unseen data. Im and Li [64] analyze how noisy feedback influences the generalization performance of preference optimization, showing that increased noise results in higher generalization risks. Specifically, standard DPO loss functions can yield biased estimates under noisy conditions. To address this

issue, rDPO [59] proposes to enhance DPO robustness against noisy annotations and improve overall training performance. Zhang et al. [63] introduce a noise-aware strategy leveraging annotator confidence and stability to identify and down-weight noisy samples during training. They also propose an adaptive reward margin, emphasizing clean samples to improve learning effectiveness. Complementary to these approaches, PerpCorrect [60] employs a data-driven method to correct noisy annotations directly in the dataset. It trains a proxy language model on both clean and noisy samples, distinguishing noise through perplexity differences to improve dataset quality. To systematically explore noise effects, Gao et al. [65] artificially inject various noise types (*e.g.*, Gaussian noise) into datasets, controlling noise intensity via hyperparameters. Their analysis highlights how noise impacts model alignment, guiding future research towards mitigating such negative effects. To address the vulnerability of DPO in noisy environments, ROPO [61] introduces a regularization term to enhance noise tolerance. Additionally, ROPO employs a robust-guided rejection sampling technique. This technique supplements the dataset with samples that contribute minimally to the loss, thereby improving the overall data quality. Kim et al. [62] propose the SPA framework, using model-generated responses and associated confidence scores to detect noise in annotations. SPA further incorporates smoothing techniques into the loss function to alleviate the noise problem. Finally, Wu et al. [66] categorize noise into two types: point noise (single annotation errors) and pairwise noise (errors between annotated pairs). While DPO naturally handles point noise well, it struggles with pairwise noise. Their proposed Dr. DPO introduces a novel loss function explicitly designed for robustness against both point and pairwise noise.

### 3.1.2 Preference Feedback

Preference feedback refers to the label signals provided by annotators regarding their preferences for different responses. It can be categorized into point-wise, pair-wise, and list-wise feedback. Point-wise feedback evaluates each response independently, assigning a score or labeling it as positive or negative. Pair-wise feedback compares two responses to determine which one is preferred, while list-wise feedback ranks multiple responses.

**(a) Point-Wise Feedback.** Point-wise feedback is the basic form of feedback. It refers to the type of feedback where individual outputs or samples are evaluated independently, rather than through comparisons with other outputs. This form of feedback is characterized by its simplicity and directness, focusing on the quality or relevance of a single response or item. The predominant methodology in RLHF [35] employs point-wise reward signals generated by reward models to optimize policy models. Similarly, KTO [67] directly maximizes the utility of model generations using loss functions based on prospect theory rather than the log-likelihood of preferences. It requires only a binary signal indicating whether an output is desirable or undesirable for a given input. Furthermore, BCO [68] builds upon the concepts introduced in KTO and explores a new approach to aligning with binary signals. While KTO focuses on optimizing human utility, BCO introduces a binary classifier framework incorporating reward shift and distribution matching that implicitly

minimizes the DPO loss. Chen et al. [72] and GPO [73] adopt explicit rewards using Noise Contrastive Alignment (NCA) and General Preference Model (GRM) respectively, and then directly optimize language model policies from point-wise preference data with rewards. However, some methods leverage implicit reward signals to refine model behaviors. To ensure that the learned implicit rewards are comparable to the ground-truth rewards, Cal-DPO [69] introduces a calibration term to the preference optimization objective, which prevents the likelihood of chosen responses from decreasing during training. ULMA [71] unifies human demonstration and point-wise preference data into a single framework and handles positive and negative samples with a hybrid objective function. Unlike them, DRO [211] adopts a simple mean-squared objective to optimize the model policy and value function jointly for a single trajectory. Additionally, AOT [70] casts the distributional preference constraint as an optimal transport problem with a convex cost function. The key idea is to minimize the violation of stochastic dominance using a smooth, convex cost function.

**(b) Pair-Wise Feedback.** Pair-wise feedback focuses on comparing pairs of data or actions to determine their relative quality or preference. Building upon the theoretical framework of RLHF, DPO implements this paradigm through the utilization of pair-wise preference data, thereby fitting an implicit reward model. Azar et al. [75] introduces a general theoretical framework to unify existing RLHF and DPO methods. The proposed Identity-Preference Optimization (IPO) directly optimizes policies from preferences without relying on reward modeling or the Bradley-Terry assumption, thereby avoiding overfitting issues observed in DPO. Subsequently, DPO-RK and DPO-R [76] integrate the Rao-Kupper and Davidson models into the DPO training objective respectively, thereby extending the capabilities of DPO by explicitly modeling ties in pairwise comparisons. BMC [77] further addresses a key limitation of the weak correlation between winning and losing responses in pairwise data. Specifically, BMC uses “Bridging” to enhance the correlation between winning and losing responses by increasing the consistency and informativeness of pairwise preference signals. However, previous attempts for aligning LLMs primarily focus on optimizing the model’s output preferences given an instruction, which struggles to effectively perceive the fine-grained constraints within complex instructions. Thus IOPO [78] extends traditional alignment methods by considering both input and output preferences to better understand the constraints within the instructions. As current methods rely heavily on paired preference data (*i.e.*, explicitly labeled preferred vs. dispreferred examples), they can be limiting in scenarios where such paired data is unavailable or insufficient. SAPO [80] addresses this issue based on the concept of self-play, which enhances data exploration and exploitation by automatically generating negative samples and integrating off-policy learning. Furthermore, PMPO [79] extends the EM algorithm to incorporate both preferred and dispreferred outcomes. By introducing the probability distribution of dis-preferred outcomes, PMPO can optimize using both types of samples, even when only negative feedback is available. Similarly, D2O [81] avoids harmful information by maximizing the discrepancy between the generated responses and the negative samples. NPO [82]



and SimNPO [83] achieve the goal of forgetting the negative impact by regulating the model’s prediction probabilities on negative datasets to be as minimal as possible, where SimNPO further eliminates the reference model bias issue inherent in NPO.

**(c) List-Wise Feedback.** List-wise feedback refers to the type of feedback where multiple outputs or responses generated by the model for a given input are evaluated collectively as a list. This approach considers the relative ranking or ordering among the outputs, rather than focusing on individual outputs in isolation. Panacea [84] reframes alignment as a Multi-Dimensional Preference Optimization (MDPO) problem and introduces a method that aims to learn the entire Pareto front to accommodate diverse user preferences. In short, Panacea is designed to adapt a single model to list-wise preferences in a Pareto-optimal manner. LiPO [85] and LIRE [86] also treat LM alignment as a list-wise ranking problem, drawing on the rich literature of Learning-To-Rank (LTR). Specifically, LiPO introduces a specific method LiPO- $\lambda$ , which leverages a list-wise ranking objective that weights each preference pair based on the difference in ranking metrics; while LIRE optimizes the response probability by calculating the exponential probability distribution and uses the reward model to directly guide the optimization process. To better capture the relative proximity within ordinal multiple responses, OPO [87] utilizes the Normalized Discounted Cumulative Gain (NDCG), a widely used ranking metric, to optimize the model’s generation probability to match the permutation of responses based on these labels. Similarly, DRPO [88] leverages NDCG as a key metric to optimize the ranking of model outputs. However, DRPO incorporates novel elements like diffNDCG and Adaptive Rank Policy Score to dynamically adjust the score margins between preferred and non-preferred responses based on their ranking positions. mDPO [232] extends preference optimization to multi-sample comparisons and introduces a framework that evaluates and optimizes the collective properties of sample groups. It not only addresses the limitations of single pair-wise methods but also provides a more robust optimization framework, especially for characteristics like diversity and bias. Furthermore, RPO [90] introduces a contrastive weighting mechanism that constructs a contrast matrix within each mini-batch to compare preferred and less-preferred responses across prompts. The weights of these comparisons are dynamically adjusted based on the semantic similarity between prompts. Additionally, TODO [91] integrates a tie ranking system into list-wise preference modeling, significantly improving the capture of nuances of human preferences, especially in the presence of noisy or inconsistent labels and frequent ties.

### 3.1.3 Preference Granularity

Preference granularity refers to the granularity of preference labels, which determines the level at which preferences are assigned to data. It can be categorized into token-level, step-level, sentence-level, and turn-level granularity, ranging from fine-grained focus on individual tokens to broader preferences over entire interaction turns.

**(a) Token-Level Granularity.** Token-level alignment operates at the character/subword unit of text generation, providing the finest-grained control over model outputs.

Theoretically, Rafailov et al. [92] demonstrate that DPO can represent any dense reward function by reparameterizing it as an optimal advantage function, which allows DPO to optimize policies in the token-level MDP effectively. TDPO [93] refines the alignment process from the sentence level to the token level and introduces forward KL divergence constraints. TDPO utilizes the Bradley-Terry model to convert sentence-level preference comparisons into a token-level reward system, which allows the model to dynamically adjust its strategy at each token generation step. Furthermore, TIS-DPO[94] estimates the importance weights of tokens based on the differences in prediction probabilities from contrastive LLMs, performing token-level importance sampling on existing data to approximate optimal distribution by assigning weights to each token based on its reward. Moreover, D<sup>2</sup>PO [99] proposes a temporal decay mechanism that dynamically adjusts the contribution of each token-level reward based on its position in the sequences. Unlike these, SparsePO [95] directly learns sparse masks during the training process and controls which tokens are more important for preferences through the sparsity of the masks, thereby achieving dynamic optimization. RTO [96] and SePO [97] first learn a token-level reward function from preference data using DPO, and then RTO optimizes PPO based on this reward signal, while SePO selects key tokens through the estimated reward function. To tackle the need for large-scale annotated data in training, EPO [98] proposes a hierarchical framework that decomposes complex tasks into manageable subgoals using separate LLMs for subgoal prediction and low-level action generation, leveraging environment feedback to automatically generate reward signals and preference data for aligning LLMs.

To conclude, token-level granularity optimizes models at individual token positions to maximize expected objectives, preserving semantic precision and capturing local syntactic dependencies. However, it increases computational complexity, as processing numerous tokens extends training time, and its sensitivity to noise means errors in a single token can affect the entire sequence. Thus, careful loss function design and regularization are essential for stability.

**(b) Step-level Granularity.** Step-level granularity focuses on the intermediate steps or stages in a process, particularly effective for complex problem-solving tasks requiring multiple intermediate steps. Step-DPO [100] and SCDPO [101] treat individual reasoning steps as the basic units for preference optimization, where preference pairs of correct and incorrect steps are generated using LLMs. Furthermore, CPO [102] and MCTS-DPO [103] first utilize more powerful inference structures to generate multiple candidate thoughts at each reasoning step following the Tree-of-Thought (ToT) and Monte Carlo Tree Search (MCTS) respectively, and construct preference pairs based on the selected and unselected intermediate steps. Then they fine-tune LLMs to generate reasoning steps preferred by ToT during inference using DPO. TPO [104] proposes a preference learning algorithm specifically designed for preference trees that have multiple branches and multi-step responses, and introduces the adaptive step reward mechanism to address the issue of small reward margins caused by shared sub-trajectories. It adjusts the reward values for each step based on semantic similarity, helping the model better distinguish

between preference pairs. RDPO [105] extends traditional preference datasets to incorporate a rationale field, which explains why a particular response is preferred. RDPO introduces rationale information into the DPO loss function by maximizing the likelihood of both the preference and the rationale, which allows the model to better understand the logic behind preferences during training. To address the challenges of sparse rewards and training instability, DAPO [106] uses a critic function to generate dense signals for policy optimization and trains the actor and critic independently to avoid instability.

To conclude, step-level alignment demonstrates unique advantages in multi-step reasoning tasks by decomposing holistic preferences into intermediate decision points. The primary strength of step-level granularity lies in its capacity to decompose complex objectives into verifiable subgoals, enhancing both interpretability and robustness. For instance, in mathematical reasoning, LLMs can receive feedback on equation derivation steps before final answers, reducing error propagation. However, this granularity still has two key challenges: first, the need for precise step segmentation, which may require domain-specific heuristics or auxiliary models to delineate reasoning boundaries; second, the risk of local optima, where over-optimization of individual steps degrades global coherence.

**(c) Sentence-level Granularity.** Sentence-level granularity aligns preferences at the complete utterance level, balancing fine-grained control and computational efficiency. This granularity, represented by the original DPO framework, operates on full response sequences as atomic units for preference comparison. MAPO [107] uses a well-trained translation model to calculate alignment scores between answers in non-dominant and dominant languages and then employs preference optimization methods to enhance reasoning consistency. EURUS [108] structures each instruction as a preference tree, containing pairs of correct and incorrect actions to facilitate preference learning. Similarly, IRPO [109] focuses on improving the reasoning capabilities of LLMs through an iterative preference optimization on constructed preference pairs such that the winning response has a higher reward than the losing response. FACTALIGN [110] proposes a fine-grained, sentence-level alignment algorithm called fKTO, which extends the KTO method to leverage fine-grained factuality assessments at the sentence level.

To conclude, the key strength of sentence-level granularity lies in its capacity to preserve holistic semantics while maintaining tractable optimization complexity. Nevertheless, we must carefully consider task requirements. While suitable for short-form generation and classification tasks, sentence-level methods may insufficiently capture fine-grained stylistic nuances or long-range dependencies critical in generation and reasoning domains.

**(d) Turn-level Granularity.** Turn-level granularity focuses on the optimization of model behavior at the level of conversational turns, which is particularly relevant for dialogue systems and interactive agents. This granularity level treats each turn of a conversation as a unit for preference alignment, allowing the LLMs to receive feedback on their responses within the context of a single turn. M-DPO [111] introduces a multi-turn direct preference learning framework to enhance the mathematical reasoning capabilities of LLMs when

integrated with external tools. It leverages feedback from code interpreters and optimizes trajectory-level preferences using signals generated by the Bradley-Terry model to improve model performance in multi-turn reasoning tasks. ETO [112] presents a novel trial-and-error learning method that optimizes LLM agents' policies by contrasting successful and failed trajectories that contain multi-turn interaction. To address the challenges of coarse granularity and training noise in previous methods, SDPO [113] optimizes specific key segments within interactions to improve multi-turn dialogues while minimizing training noise. Specifically, it extracts key segments from the positive sessions that contribute to higher goal and relationship scores and pairs them with corresponding segments from the negative sessions to calculate an adapted DPO loss. Similarly, AgentQ [114] combines MCTS with self-critique mechanisms to provide process-level supervision by ranking actions, and then iterative fine-tuning using DPO. This approach enables LLMs to effectively learn from both successful and unsuccessful trajectories, enhancing their generalization and decision-making capabilities in complex, multi-turn reasoning tasks within interactive environments. DMPO [115] enhances the existing DPO method by replacing the policy constraint with a State-Action Occupancy Measure (SAOM) constraint and incorporating length normalization into the Bradley-Terry model, effectively addressing challenges in multi-turn scenarios. Compared to traditional policy constraints, SAOM constraints better guide the agent to select actions that align with expert trajectories, especially in unexplored states, thereby reducing compounding errors.

To conclude, turn-level alignment offers critical advantages for interactive systems by optimizing contextually grounded responses while preserving conversational flow. However, in multi-turn dialogue tasks, the turn-level granularity may introduce additional training noise. For example, some correct turns in negative samples may be mistakenly treated as incorrect turns in the loss calculation. Additionally, since each turn needs to be processed independently, this can lead to reduced training efficiency.

## 3.2 Learning Framework of DPO

The learning framework of DPO focuses on how the language model policy learns from preference data. In this section, we present an overview of the learning framework in DPO, as shown in Fig. 3, which encompasses the learning paradigm and the learning objectives.

### 3.2.1 Learning Paradigm

The learning paradigm in DPO determines how preference data is acquired during model training and falls into three distinct categories: offline learning, where the model learns from pre-collected preference datasets; online Learning, where the model updates based on newly generated data; and active Learning, where the model selectively queries annotators obtain preference data.

**(a) Offline Learning.** The original DPO framework [74] itself is an offline learning paradigm, where the model learns from a static, pre-collected dataset of preference pairs. Recent research has explored different approaches to merging preference optimization and supervised fine-tuning



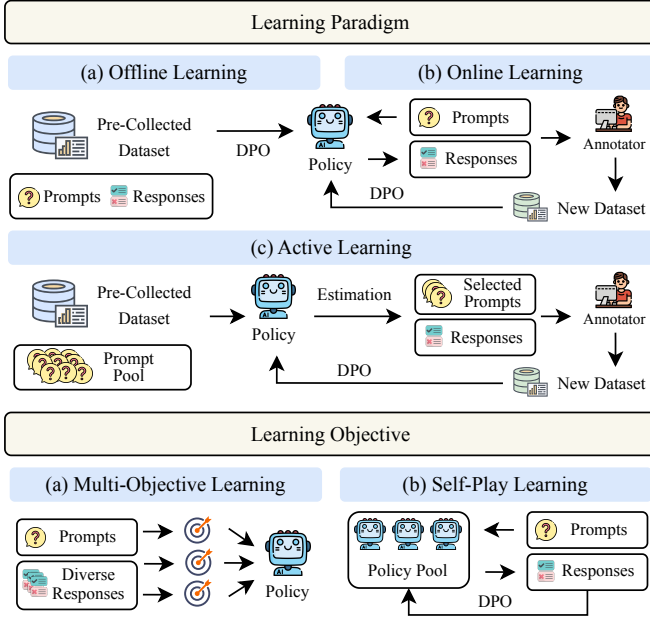


Fig. 3: An overview of DPO learning framework.

into a single training phase [190]. CPO [116] incorporates a behavior cloning regularizer through KL divergence minimization between the model and preferred data distribution, which effectively combines into adding a negative log-likelihood term on preferred data alongside the contrastive preference loss. Taking a more direct approach, ORPO [117] proposes a monolithic framework that directly augments the standard negative log-likelihood loss with an odds ratio term comparing chosen and rejected responses, eliminating the need for a separate reference policy while preserving SFT’s domain adaptation capabilities. ULMA [71] proposes a hybrid method that applies standard SFT loss on positive samples while using a ranking-based DPO loss on negative samples. PAFT [118] introduces a parallel training paradigm where SFT and preference alignment are performed concurrently on the same pre-trained model and then merged using parameter fusion techniques, avoiding the sequential pipeline that can lead to catastrophic forgetting.

Several advances explore curriculum learning strategies to enhance DPO performance and training efficiency. Curri-DPO [53] introduces curriculum learning by ordering multiple preference pairs from easy to hard based on the rating difference between chosen and rejected responses, where pairs with larger rating gaps are presented first, followed by progressively more challenging pairs with smaller rating differences. sDPO [55] implements curriculum learning by partitioning preference datasets into sequential chunks measured by reward accuracy and applying them incrementally.

To avoid substantial computational and data annotation costs for preference alignment, fine-tuning-free alignment methods have gained popularity. Linear Alignment [119] works by directly estimating the optimal policy through a one-step update to the output distribution during inference without requiring parameter tuning or feedback data. ICDPO [120] reinterprets DPO’s reward-policy relationship to create a fine-tuning-free alignment method that harnesses in-context learning, treating models before and after demonstration exposure as amateur and expert policies, respectively,

then computing their log probability ratio to score and rank candidate responses.

**(b) Online Learning.** DPO faces significant limitations when relying solely on static, pre-collected preference datasets. These datasets, generated by different models, cause a distribution shift that leads to ineffective off-policy learning as the model evolves [145, 152]. By contrast, online DPO employs an iterative framework that continuously updates the policy with real-time feedback, ensuring on-policy learning and reducing misalignment [143, 144, 233].

As online DPO progresses, researchers have introduced more flexible frameworks to tackle key challenges. For instance, Yuan et al. [123] proposed a self-rewarding language model: the model generates prompts and responses, then serves as its own judge via LLM-as-a-Judge prompting, scoring on a 5-point scale. OAIF [121] uses an LLM as an online annotator for real-time feedback, and OFS-DPO [122] addresses catastrophic forgetting by using two Low-Rank Adaptive (LoRA) modules with different optimization speeds. BPO [124] constructs a dynamic trust region around the behavior LLM, adjusting it as preference data is collected, unlike methods that rely on fixed reference models. Furthermore, researchers have refined sampling strategies for online DPO. RSO [126] and RS-DPO [125] employ rejection sampling based on reward gaps. ROPO [61] recovers useful information from discarded queries via robustness-guided rejection sampling. Shi et al. [127] introduced DPO-Mix-R and DPO-Mix-P, demonstrating faster convergence by mixing online samplers with uniform samplers. OPTUNE [128] selectively regenerates responses with low reward scores while reusing high-reward responses. Iterative RPO [109] and DPO-ST [129] enhance CoT reasoning by selecting correct and incorrect answers to form preference pairs at each iteration. Xie et al. [103] used MCTS to collect preference data during training. Researchers have also explored advanced optimization techniques. APO [130] incorporates momentum-based acceleration, using an extrapolation step between the current and previous policies to update the policy. Xiong et al. [131] proposed a two-agent, non-symmetric online DPO framework with a main agent for optimal policy learning and an enhancer agent for exploration. COMAL [132] formulates alignment as a two-player zero-sum game, updating its policy toward a regularized Nash equilibrium in each iteration. PCO [133] iteratively trains the model on preference data with pairwise cringe Loss.

Recent efforts push for greater autonomy by letting models generate their own feedback [62]. SeRA [134] introduces a self-reviewed preference bootstrapping method, using an implicit reward margin to select informative pairs, and employs an ensemble reward approach across iterations. CREAM [135] mitigates self-improving biases by applying a consistency regularization on the preference rankings of consecutive iterations. D2PO [136] combines human-labeled gold data with concurrently updated, discriminator-labeled data. DLMA [137] uses contrastive prompts to compute self-reward scores via log ratio differences, then integrates these scores directly into the DPO objective. Addressing exploration and uncertainty in online DPO has also been a focus [234]. XPO [138] encourages exploration by adding a bonus for responses outside the initial policy’s support, and SELM [139] uses an optimism term in reward fitting to

actively seek high-reward responses. ETO [112] alternates exploration and training phases to collect failure trajectories, while VPO [140] applies optimism by regularizing the reward model to favor higher-value responses. Xiong et al. [111] extended DPO from single-turn to multi-turn tasks, balancing KL-regularized and non-regularized objectives, and COPO [141] incorporates a count-based bonus to encourage novel responses with low visitation counts.

Finally, a growing body of work aims to merge online and offline techniques. HyPO [142] uses offline preference data for DPO training while regularizing via online data. MPO [47] combines the strengths of DPO and PPO in a two-stage process: it first trains DPO on an easier dataset, then uses this model as a reference for PPO training on more challenging samples.

**(c) Active Learning.** Active learning in DPO is a strategic approach that aims to reduce the annotation cost and improve sample efficiency by selectively querying annotators for the most informative preference examples. Unlike offline learning that uses a fixed dataset or online learning that generates new data continuously, active learning intelligently selects which data points should be labeled based on model uncertainty or other informativeness criteria.

Muldrew et al. [146] introduced APL, an iterative data acquisition and fine-tuning loop in which batches of prompt/completion pairs are strategically selected using acquisition functions: a predictive entropy-based approach to measure model uncertainty for prompts and a preference certainty measure based on the implicit Bradley-Terry model for completion pairs in DPO. Unlike two-step selection processes in APL that separately select uncertain input prompts and corresponding completions, divAPO [147] integrates both stages into a single selection phase. divAPO maximizes the preference model certainty by simultaneously evaluating the informativeness of input prompts and completion pairs, while also considering the data distribution of the input prompts. Ji et al. [148] proposed ADPO, which selectively queries human preferences only for responses where the model exhibits high uncertainty while using pseudo-labels for confident cases. Das et al. [149] also employed active learning on RLHF, which actively selects the context-action pairs that maximize exploration and minimize uncertainty in the reward model.

### 3.2.2 Learning Objective

In what follows, we present the learning objective in DPO, which determines how the model policy is optimized based on preference data. We first discuss multi-objective learning in DPO, which aims to optimize multiple objectives simultaneously. Then, we introduce self-play learning, which leverages self-generated data for preference alignment.

**(a) Multi-Objective Learning.** Multi-objective learning in DPO addresses the challenge of simultaneously optimizing the language model for multiple, potentially competing preference dimensions, such as helpfulness, harmlessness, and truthfulness. This approach aims to find a balanced policy that satisfies multiple human values rather than optimizing for a single objective, which more closely mirrors the complexity of real-world human preferences.

MODPO [150] achieves the sequential optimization of multiple preference objectives by incorporating language

modeling directly into reward modeling, using a margin-based loss to maintain performance on previously optimized dimensions. SPO [151] takes a similar iterative constrained optimization approach, optimizing each preference dimension while preventing the degradation of prior alignments through regularization terms. MOSLIM [152] takes a different approach by introducing a multi-head classification reward model that assigns different preference dimensions to separate classification heads, enabling simultaneous optimization of multiple preferences without requiring multiple reward or policy models. HPO [153] incorporates auxiliary objectives through offline RL, where the model uses a weighted maximum likelihood objective that combines a preference alignment term with an advantage-weighted term for maximizing arbitrary auxiliary rewards like readability and safety. CPO [154] introduces explicit preference tokens during training that specify desired scores for different objectives, transforming the multi-objective optimization into a conditional optimization problem. DRDO [155] simultaneously models rewards and preferences through a combination of reward distillation and a contrastive log-unlikelihood term in its loss function.

**(b) Self-Play Learning.** Self-play learning in DPO represents an approach where the language model interacts with itself or its previous iterations to generate its own preference data for training, reducing or eliminating the need for human annotations [139, 164]. This method enables continuous self-improvement by leveraging the model’s own judgment capabilities to identify and learn from better responses, creating a form of autonomous preference learning.

SPIN [156] involves a self-play mechanism where the LLM generates synthetic data from its prior iterations, then fine-tunes itself to distinguish these self-generated responses from those of human-annotated data. The method resembles a two-player game, where the model’s current iteration tries to improve its responses to better match the target distribution, while the previous iteration attempts to generate responses as close to human data as possible. SPPO [157] treats LLM alignment as a constant-sum two-player game and iteratively refines itself by competing against its previous iteration. Instead of maintaining two competing policies or a reward model, SPO [158] uses a single policy to sample multiple trajectories and uses the proportion of wins in pairwise comparisons as the reward signal. BoNBon [159] Alignment likewise relies on sampling responses from a base model, but it selects the best ones among  $n$  candidates and fine-tunes itself to approximate that best-of- $n$  distribution.

Some works approach the alignment problem by leveraging Nash equilibrium [132]. Nash-MD [160] learns a preference model from pairwise human feedback and then computes a Nash equilibrium policy that consistently produces preferred responses. Its self-play approach updates the policy by having it compete against itself (or a slight variant of itself) under the learned preference model, which measures how often one response is preferred to another. DNO [161] extends this concept by implementing a batched on-policy algorithm where the current policy generates multiple outputs that are compared both against each other and against a teacher model’s outputs. IPO-MD [162] combines the strengths of IPO and Nash-MD, where the model generates data using a mixture policy between the online and reference

policies, and uses a preference model to annotate pairs of generations, making the optimization equivalent to finding a Nash equilibrium through self-play. SRPO [163] modifies Nash-MD by introducing a self-improvement policy that refines model outputs through iterative revisions, enabling offline optimization without a learned reward function.

### 3.3 Constraint Mechanism of DPO

The constraint mechanism of DPO derives from its reformulation of RLHF, which includes a KL divergence constraint between the current policy and a reference policy. As shown in Fig. 4, we re-examine the constraint mechanism of DPO from the perspective of the reference model and different divergence constraints. We also explore various DPO variants with different safety constraints.

#### 3.3.1 Reference Model

The reference model in DPO functions as an anchor to ensure policy updates remain within a controlled range, preventing excessive deviation from initial behaviors. Typically, the reference model is initialized using the SFT model that serves as the starting point for preference optimization. The choice of reference model significantly impacts optimization dynamics. A static reference model ensures stable training but may limit adaptability. In the following subsections, we introduce two advanced approaches: reference-free DPO eliminates reliance on the reference model, while dynamic-reference DPO employs an evolving reference model.

**(a) Reference-Free DPO.** To reduce the computational and memory costs associated with a reference model, many algorithms have explored training modes that do not require loading the reference model. Xu et al. [116] replaces the reference model with a uniform prior distribution, adding an SFT loss term on preferred data to maintain consistency with the desired behavior. ORPO [117] integrates an odds ratio-based penalty with traditional SFT loss, increasing the probability of preferred responses while decreasing undesirable ones, thereby enabling single-stage training without a separate reference model. SimPO [166] directly uses the average log probability as implicit rewards. This removes the requirement for a separate reference model, significantly improving computational and memory efficiency. SimPER [167] also directly optimizes reverse perplexity for preferred versus rejected responses, creating a preference optimization approach that does not require a separate reference model, thus simplifying training. Despite these advancements, [168] argue that a reference model remains crucial. They compared two reference-free variants using posterior probabilities and likelihood functions as rewards, respectively, and found the original DPO consistently outperformed both. Their results indicate that a strong, well-aligned reference policy can significantly enhance DPO performance.

**(b) Dynamic-Reference DPO.** Offline DPO methods often suffer from reward over-optimization, meaning that as the trained model deviates from the reference model, the quality of generated samples tends to degrade. To address this issue, Gorbатовski et al. [165] proposed dynamically updating the reference model using the current model parameters during training, preventing excessive divergence and maintaining high-quality outputs. Curri-DPO [53] and sDPO [55] adopt

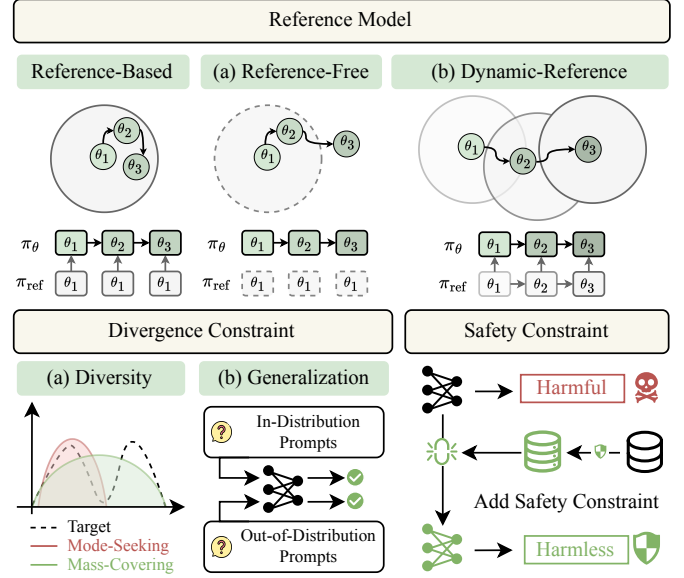


Fig. 4: An overview of DPO constraint mechanism.

curriculum learning by sorting data samples from simpler to more complex based on predefined metrics. At each training iteration, the model from the previous step serves as the updated reference model to provide constraints, facilitating progressive learning. Similarly, MPO [47] partitions datasets according to task difficulty, employing a two-stage training procedure. The model trained in the initial stage serves as the reference for the subsequent stage. Additionally, M-DPO [89] compares the performance of a fixed reference model versus a dynamic reference model, finding that the latter yields superior results.

#### 3.3.2 Divergence Constraint

Divergence constraints in DPO play a crucial role in constraining model optimization, balancing alignment performance and model stability. In the following subsections, we introduce two modifications to the divergence constraint: one for enhancing diversity and the other for improving generalization.

**(a) Diversity.** Standard DPO typically uses reverse KL divergence equivalent to RLHF. However, the mode-seeking nature of reverse KL divergence reduces the diversity of the generated outputs. To overcome this limitation, f-DPO [169] explores various divergences, including forward KL divergence, reverse KL divergence, Jensen-Shannon divergence, and  $\alpha$ -divergence, to achieve a better trade-off between alignment performance and diversity. Slocum et al. [170] further proposes splitting the KL divergence term into entropy and cross-entropy terms. This decoupling allows independent control of generation diversity and closeness to the reference model, preserving output diversity without degrading overall model quality.

**(b) Generalization.** Over-optimization in DPO can negatively impact generalization, causing reduced performance on inputs outside the training distribution. To mitigate this, Huang et al. [178] introduce  $\chi^2$ -divergence as a more aggressive form of regularization compared to KL divergence, alleviating the over-optimization problem. DPO-Kernels [171] employs data-driven methods to select optimal kernel-divergence pairs dynamically, improving task adaptability



and robustness. FlipGuard [172] introduces a customized reward characterization to monitor model performance. If performance drops relative to earlier versions, FlipGuard constrains the model’s updates to ensure alignment with previous stable behavior. FPO [173] leverages the feature-level constraints using Sparse Autoencoders (SAEs) to improve computational efficiency and training stability. SPO [176] integrates a natural preference loss with a KL divergence-based regularization term computed over the entire model output distribution. By adjusting this divergence term, SPO prevents unwanted shifts beyond the preference dataset, ensuring stable alignment. EXO [175] argues that minimizing the forward KL divergence in DPO introduces bias when approximating the optimal policy. They establish a generalized alignment objective and reveal the equivalence between maximizing KL regularization rewards and minimizing the reverse KL divergence relative to the optimal policy. QDPO [177] utilizes divergence between the quantized model and the full-precision model for preference optimization, effectively addressing the token-flipping issue. Token-flipping refers to the phenomenon where quantization errors skew token distributions, leading to incorrect token selection. GPO [174] constructs a framework that unifies different DPO-related algorithms through theoretical derivations, enabling a deeper understanding of the regularization mechanisms in the DPO family of algorithms.

### 3.3.3 Safety Constraint

Safety constraints in DPO aim to prevent LLMs from generating harmful, biased, or unethical outputs. However, traditional alignment algorithms often fail to address safety concerns. To enhance the safety alignment, recent studies have introduced several specialized mechanisms based on DPO. SafeDPO [179] introduces a streamlined approach for safety alignment by implicitly optimizing safety objectives within a single stage of policy learning. SACPO [180] addresses safety constraints by explicitly formulating language model alignment as a constrained optimization problem, using DPO to optimize the model under safety constraints. Zhang et al. [184] propose creating a backtracking preference dataset that identifies and reverses unsafe outputs, enhancing the safety and robustness of the model. C-DPO [181] integrates dual gradient descent into DPO to balance safety and utility efficiently. This approach achieves a robust trade-off between helpfulness and harmlessness, offering explicit safety guarantees. ADPO [182] introduces adversarial techniques into DPO. It specifically trains models to reduce the probability of unsafe outputs by deliberately generating harmful responses using controlled toxic tokens. Finally, Lee et al. [183] explore the internal mechanisms through which DPO reduces harmful outputs. Their findings suggest that DPO does not remove harmful behaviors learned during pre-training but instead teaches models to bypass or suppress these behaviors. This insight helps explain certain safety vulnerabilities like jailbreaks.

## 3.4 Model Property of DPO

DPO has shown great promise in aligning LLMs with human preferences by directly optimizing model outputs based on preference data. During this process, the underlying models

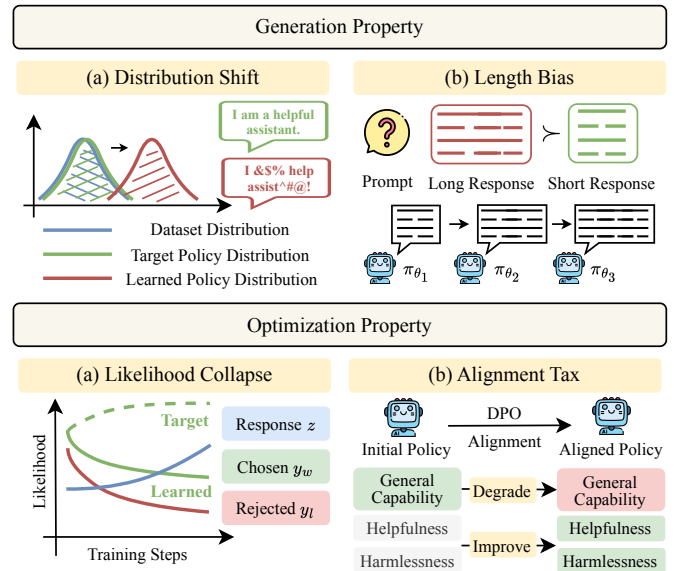


Fig. 5: An overview of DPO model property.

exhibit certain properties that are crucial for understanding their behavior and effectiveness. These properties can be broadly categorized into two aspects: the generation property and the optimization property, as shown in Fig. 5. In the following sections, we explore these two properties in more detail, analyzing their implications for model alignment.

### 3.4.1 Generation Property

The generation property of DPO primarily concerns issues related to distribution shifts and length biases. DPO is sensitive to distribution shifts between the base model outputs and the preference data, which may reduce diversity and generalization. Additionally, DPO has a tendency to favor longer responses, a phenomenon known as verbosity, which can negatively impact performance and user experience.

**(a) Distribution Shift.** In RLHF, the reward model is trained on a static set of preference data collected offline. During fine-tuning, the generated responses often differ from this original training data, resulting in a distribution shift. This shift can cause inaccurate reward predictions and lead to over-optimization. The implicit reward model in DPO also suffers from this distribution shift issue. Moreover, Lin et al. [188] have shown that the implicit reward model in DPO performs poorly on Out-Of-Distribution (OOD) data compared to explicit reward models. Experimental results indicate that DPO can transfer probability mass to the high-reward response regions covered by the preference data, but it may also cause the distribution of responses generated by the model to deviate significantly from that of the reference policy, resulting in responses that do not meet expectations [189]. To address these problems, many researchers are now exploring online DPO approaches [109, 121, 122, 125], aiming to mitigate OOD by continuously updating preference data during training.

Existing DPO methods also face significant limitations due to their dependence on specific training tasks. Their optimal solutions lack robustness when applied to OOD tasks. Thus, SRPO [163] reframes alignment as a self-improvement process, which optimizes a self-improvement policy and a generative policy using a min-max objective, ensuring

robustness by making the solution independent of training tasks. Zhang et al. [139] also identify notable issues in DPO when handling OOD tasks. First, DPO tends to overly favor novel content it has not seen during training. Second, it easily gets stuck in local optima, limiting exploration. To address these problems, they propose Self-Exploring Language Models (SELM), incorporating an optimism term to encourage broader exploration of new responses.

Another significant challenge of DPO is preference drift, where human preferences evolve, changing data distributions over time. Traditional DPO algorithms typically overlook such temporal shifts, mistakenly interpreting them as noise. To address this, NS-DPO [185] propose to assign higher weights to recent data, allowing models to better adjust to evolving preferences.

**(b) Length Bias.** Length bias in DPO refers to the tendency of model-generated outputs to become excessively long during training. This issue is similar to the length bias observed in RLHF [197] and is particularly pronounced in DPO. Length bias affects response quality and overall model performance. To mitigate this issue, researchers have developed several solutions, which can be categorized into three main approaches: length regularization, length normalization, and length sampling.

Length regularization is a common approach to controlling length bias in DPO. By introducing regularization terms into the objective function, the model can constrain response length and reduce verbosity, thereby alleviating the length bias problem. R-DPO [191] introduces a length-based penalty term to the DPO objective function, explicitly discouraging verbosity. D<sup>2</sup>PO [99] introduces a dynamic weighting mechanism by incorporating a temporal decay factor. Unlike previous methods that apply uniform reward contributions across sequences, D<sup>2</sup>PO adjusts the influence of each reward based on its position in the response. Higher weights are assigned to rewards associated with earlier tokens, as they are more critical for model alignment, while later rewards gradually receive lower weights. This adaptive approach prevents overfitting to less relevant tokens, thereby addressing length bias in DPO.

Length normalization aims to eliminate the loss bias caused by response length differences, allowing the model to evaluate texts of varying lengths more fairly. This approach prevents the model from developing an unreasonable preference for either long or short responses [198]. RRHF [190] and SimPO [166] first propose to apply length normalization to responses, ensuring that the loss remains unaffected by response length. LN-DPO [194] further integrates SimPO-like length normalization into DPO, demonstrating that this approach enhances response quality while mitigating verbosity. LD-DPO [195] achieves length desensitization by reparameterizing the likelihood in DPO. Specifically, it decomposes the likelihood of the longer response in a preference pair into the product of the likelihood of the public-length portion and the likelihood of the excessive portion. It then introduces a hyperparameter to mitigate the verbosity preference. This adjustment smooths the relationship between likelihood and response length, reducing its impact on optimization. For multi-turn dialogue tasks, DMPO [115] introduces length normalization for the number of turns in multi-turn preference optimization.

An alternative approach to controlling length bias in DPO is through sampling-based methods. SamPO [192] introduces a down-sampling method to compute regularized KL divergences. By balancing token-level probability distributions between preferred and rejected responses, SamPO reduces length bias in DPO training. Yuan et al. [193] propose Length-Instruction Fine-Tuning (LIFT), a method to improve instruction-following models’ ability to adhere to length constraints by augmenting existing training data with explicit length instructions and using DPO for training. This enables the model to generalize across prompts requiring different response lengths. For long-context tasks, LongPO [196] enables short-context LLMs to self-evolve for long-context tasks by learning from self-generated short-to-long preference data, which includes paired responses for long-context inputs and their compressed short-context counterparts. LongPO incorporates a short-to-long KL constraint to prevent degradation of short-context performance during long-context alignment, achieving strong performance on both short- and long-context tasks.

### 3.4.2 Optimization Property

The optimization property of DPO involves likelihood collapse and alignment tax. While DPO aims to increase the likelihood of preferred responses and decrease dispreferred ones, the actual optimization process does not explicitly enforce this balance. Moreover, alignment improvements often come at the cost of the original capabilities of LLMs, known as alignment tax.

**(a) Likelihood Collapse.** Likelihood collapse refers to the unintended reduction in the likelihood of both preferred and dispreferred responses during DPO training [92]. This phenomenon can lead to unintentional unalignment, where the model’s outputs deviate from human preferences, potentially producing undesirable or harmful responses. This phenomenon is also referred to as likelihood displacement in prior studies [204]. Additionally, the gradients associated with increasing the likelihood of preferred responses and decreasing that of dispreferred responses can become entangled, hindering effective learning. This entanglement complicates the optimization process, making it challenging to achieve the desired alignment [203]. Theoretical analyses have further elucidated the underlying causes of likelihood collapse. In particular, Feng et al. [202] developed an analytical framework grounded in field theory. Their analysis of the gradient vector field of the DPO loss function revealed that the loss function decreases the probability of generating human-disliked data at a faster rate than it increases the probability of generating human-liked data.

Several strategies have been proposed to address likelihood collapse. Pal et al. [200] introduce DPO-Positive (DPOP), which adds a penalty term to maintain a high log-likelihood for preferred examples. Similarly, LLaMA [235] augments DPO training with a negative log-likelihood term to stabilize training and preserve the log-likelihood of chosen responses [109]. Flex-DPO [201] adaptively adjusts parameters to slow the decline in the likelihood of dispreferred responses and balance gradients for both chosen and rejected outputs. D’Oosterlinck et al. [199] propose Anchored Preference Optimization (APO), which provides fine-grained control over probability updates: APO-zero increases the

TABLE 1: An overview of datasets (upper row) and benchmarks (lower row) for DPO.

Dataset	Task Description	Data Size (Training & Test)		Data Source	Data Structure	Evaluation Metric
UltraFeedback [237]	Instruction-Following, Helpful	64K	& -	AI	List	-
SafeRLHF [238]	Harmless, Helpful	73.9K	& 8.21K	Human&AI	Pair	-
HelpSteer [239]	Helpful	35.3K	& 1.8K	Human	Point	-
PRM800K [240]	Mathematical Reasoning	800K	& -	Human	Point	-
SHP-2 [241]	Q&A From Reddit	3600K	& 241K	Human	Pair	-
Nectar [242]	Conversations	183K	& -	AI	List	-
OpenOrca [243]	Conversations	2940K	& -	AI	Sample	-
Capybara [244]	Multi-Turn Conversations	16K	& -	Human&AI	Sample	-
Step-DPO [100]	Mathematical Reasoning	10.8K	& -	Human&AI	Pair	-
BeaverTails [245]	Harmless, Helpful	330K	& 36K	Human&AI	Point	-
IMDb [246]	Movie Reviews	25K	& 25K	Human	Sample	Accuracy
Reddit TL;DR [247]	Summarization	1330K	& -	Human	Sample	Win Rate
Anthropic-HH [248]	Harmless, Helpful	161K	& 8.55K	AI	Pair	Win Rate
GSM8K [249]	Mathematical Reasoning	7.47K	& 1.32K	Human	Sample	Accuracy
AlpacaEval2 [250]	Automatic Evaluation	-	& 0.8K	AI	Sample	Win Rate
MT-Bench [251]	Multi-Turn Question	-	& 3.3K	Human	Pair	Win Rate
AdvBench [252]	Harmful Behaviors	-	& 0.5K	Human	Sample	Attack Success
Arena-Hard [253]	Updating Evaluation	-	& 0.5K	AI	Sample	Win Rate
TruthfulQA [254]	Truthful	-	& 0.8K	Human	Pair	Accuracy
IFEval [255]	Instruction-Following	-	& 0.5K	Human	Sample	Accuracy
BBH [256]	Multistep Reasoning	-	& 23 Tasks	Human	Sample	Accuracy
MATH [257]	Mathematical Reasoning	7.5K	& 5K	Human	Sample	Accuracy
GPQA [258]	Biology, Physics, and Chemistry	-	& 0.45K	Human	Sample	Accuracy
MUSR [259]	Multistep Reasoning	-	& 0.76K	AI	Sample	Accuracy
MMLU-Pro [260]	Language Understanding	-	& 12K	Human&AI	Sample	Accuracy

probability of winning outputs and decreases that of losing outputs, whereas APO-down decreases both, but with a stronger decline for losing outputs.

Another notable challenge related to likelihood collapse is likelihood over-optimization, where the performance of a model on a proxy metric (such as its own likelihood estimates) improves, while its true performance does not. Zhang and Ranganath [236] show that reductions in the likelihood loss of DPO do not necessarily translate into higher win rates. Similarly, Shi et al. [205] further investigates the problem of likelihood over-optimization in DPO, demonstrating that higher completion likelihoods do not necessarily correlate with better model performance and may even degrade it. This study identifies key indicators of over-optimization and highlights the need to balance likelihood optimization with output diversity. e-DPO [187] also shows that DPO can lead to degenerate policies due to overfitting, and proposes a solution using reward model distillation to regularize the implicit reward of the language model. The method trains the language model to match the probability distribution induced by a reward model and introduces a pessimistic extension to handle uncertainty in the reward model, thereby improving the robustness of DPO.

**(b) Alignment Tax.** Alignment tax refers to the unintended consequence where improving a model’s preference alignment degrades its general capabilities acquired during pretraining [206]. Thakkar et al. [207] demonstrate the sensitivity of DPO to training data composition, showing significantly worse performance degradation than SFT when using mixed-preference datasets. Furthermore, Chen et al. [209] identify that DPO struggles with optimizing ranking tasks. While DPO improves ranking accuracy, it disproportionately harms generative capabilities. Pentyla et al. [118] also observes capability forgetting during sequential training, where DPO objectives conflict with previously learned SFT patterns. To address this, researchers propose model merging strategies that balance alignment and performance.

PAFT [118] separately trains SFT and DPO objectives on a pretrained model using distinct datasets, then merges the parameters through weighted averaging. Additionally, Lu et al. [208] proposes online merging optimizers, which integrate model merging into each optimization step of DPO to balance human preferences and basic capabilities. By merging gradients with parameter differences between SFT and pretrained models, these optimizers effectively enhance alignment while mitigating alignment tax.

## 4 BENCHMARKS AND ANALYSIS

In this section, we provide a comprehensive overview of existing benchmarks and evaluation for DPO methods. We first introduce the key datasets and benchmarks used to train or evaluate DPO models. We then present a comparative analysis of the performance of different DPO methods on these benchmarks, highlighting their strengths and limitations.

### 4.1 Datasets and Benchmarks

A diverse range of datasets and benchmarks has been specifically curated to facilitate research in DPO. Table 1 summarizes these datasets and benchmarks, highlighting their task descriptions, dataset sizes, data sources, data structures, and evaluation metrics. These datasets and benchmarks span a broad range of tasks, such as harmlessness and helpfulness evaluation and mathematical reasoning. They also exhibit significant diversity in scale, ranging from smaller, specialized datasets to large-scale collections such as SHP-2, which contains over 3.6 million samples. Additionally, datasets differ in their sources: some rely purely on human annotations, others on AI-generated content, and many adopt a hybrid approach combining human and AI-generated data. The data structures employed across these datasets include single-sample without preference label, point-wise annotations, pair-wise comparisons, and list-wise comparisons. Common evaluation metrics include accuracy



TABLE 2: Experimental results of different DPO variants on Open LLM Leaderboard. The underline indicates the best performance.

Model	Mistral-7B-Base							LLaMA-3-8B-Base						
	IFEval	BBH	MATH	GPQA	MUSR	MMLU-Pro	AVERAGE	IFEval	BBH	MATH	GPQA	MUSR	MMLU-Pro	AVERAGE
SFT	3.4	41.1	9.2	28.8	42.0	27.7	25.4	29.0	46.3	15.3	28.6	41.3	31.0	31.9
RRHF [190]	10.0	40.6	1.7	26.4	46.3	26.1	25.2	31.0	46.8	13.9	31.4	36.8	30.5	31.7
SLiC-HF [230]	11.0	44.0	9.9	29.2	42.6	28.1	27.5	41.7	49.5	17.5	30.4	39.7	31.7	35.1
DPO [74]	11.1	43.7	7.1	28.5	43.8	26.7	26.8	34.3	48.2	17.2	31.9	40.1	31.5	33.9
IPO [75]	9.4	42.8	9.7	29.7	39.7	27.8	26.5	35.3	49.0	15.9	32.8	41.4	31.9	34.4
CPO [116]	8.0	42.7	9.6	28.9	42.1	27.3	26.4	32.4	46.9	16.8	30.6	39.1	31.8	32.9
KTO [67]	12.9	43.7	12.0	28.9	46.1	28.3	28.6	40.2	48.3	18.0	31.0	40.1	31.1	34.8
ORPO [117]	28.4	46.4	13.5	30.2	41.4	29.5	31.6	40.0	49.1	16.8	30.7	38.4	32.0	34.5
R-DPO [191]	10.0	43.0	7.6	28.7	39.3	27.2	26.0	36.4	48.8	17.2	31.6	40.6	31.5	34.4
SimPO [166]	11.1	43.1	8.4	28.9	39.5	27.2	26.4	40.8	48.6	15.8	31.0	40.5	31.8	34.7

Model	Mistral-7B-Instruct							LLaMA-3-8B-Instruct						
	IFEval	BBH	MATH	GPQA	MUSR	MMLU-Pro	AVERAGE	IFEval	BBH	MATH	GPQA	MUSR	MMLU-Pro	AVERAGE
SFT	48.4	46.2	10.9	29.1	47.6	27.1	34.9	50.7	49.3	26.9	31.0	37.9	35.7	38.6
RRHF [190]	45.2	45.3	10.1	28.5	44.2	26.2	33.3	51.3	49.3	27.2	29.6	39.5	35.3	38.7
SLiC-HF [230]	39.4	46.2	11.4	28.7	49.0	26.8	33.6	41.6	50.9	26.3	31.3	39.2	35.3	37.4
DPO [74]	49.0	45.6	11.0	26.9	46.1	26.8	34.2	48.9	50.1	25.8	29.4	38.7	36.0	38.2
IPO [75]	42.6	45.3	11.8	27.8	49.3	27.2	34.0	50.4	49.5	26.3	29.6	37.9	35.7	38.2
CPO [116]	38.8	46.0	10.1	28.5	48.4	26.9	33.1	50.6	49.1	26.8	31.3	38.1	35.8	38.6
KTO [67]	46.2	45.7	10.9	27.8	46.0	27.3	34.0	43.1	50.1	26.3	31.2	38.1	35.0	37.3
ORPO [117]	37.6	45.1	11.2	28.2	46.9	26.5	32.6	43.0	50.6	26.9	29.3	39.1	35.1	37.3
R-DPO [191]	46.8	45.9	9.9	28.7	46.2	27.6	34.2	50.9	50.3	25.3	29.8	39.0	35.7	38.5
SimPO [166]	45.4	45.9	10.4	28.3	45.0	27.1	33.7	48.8	49.2	25.0	29.3	39.2	35.1	37.8

(for tasks like mathematical reasoning found in GSM8K and MATH), win rates derived from pairwise comparisons (such as MT-Bench and Anthropic-HH), and attack success rates used for assessing adversarial robustness (AdvBench).

## 4.2 Results

To demonstrate the effectiveness of different DPO variants, we conduct experiments on the Open LLM Leaderboard. We compare different DPO variants using Mistral-7B-Base, Mistral-7B-Instruct [261], LLaMA-3-8B-Base, and LLaMA-3-8B-Instruct [235] as starting points. The overall experimental setup follows Meng et al. [166], ensuring a reproducible evaluation of different DPO variants. For Mistral-7B-Base and LLaMA-3-8B-Base, the SFT models are trained based on the UltraChat-200k dataset [262], and subsequently applied different DPO variants on the SFT models using the UltraFeedback dataset [237]. For Mistral-7B-Instruct and LLaMA-3-8B-Instruct, which have already undergone instruction-tuning, the preference dataset is regenerated by collecting responses from the SFT models using prompts from the UltraFeedback dataset [237].

The experimental results, as summarized in Table 2, highlight the performance of different DPO variants across various benchmarks. For the Mistral-7B-Base and LLaMA-3-8B-Base models, ORPO consistently achieves the highest average scores, indicating its effectiveness in aligning models with human preferences. Notably, ORPO outperforms other methods on IFEval, BBH, and MATH, demonstrating its superiority in instruction-following and mathematical reasoning tasks. Meanwhile, SLiC-HF and KTO also achieve competitive results, particularly in BBH and GPQA, suggesting that these methods effectively leverage preference data for enhanced performance. For the Mistral-7B-Instruct and LLaMA-3-8B-Instruct models, the improvements across different DPO variants are more nuanced. While DPO and R-DPO show strong performance in IFEval and MMLU-Pro, IPO and CPO demonstrate robustness in handling complex reasoning tasks like MATH and GPQA. Overall, the results indicate that different DPO variants exhibit varying strengths across benchmarks, with some methods excelling in base models while others are more effective for instruct models.

## 5 APPLICATIONS

In this section, we discuss the applications of DPO in various domains, including different LLM-based applications, diffusion models, and multi-modal LLMs. We provide an overview of the key challenges and opportunities in each domain and highlight the potential impact of DPO on real-world applications.

### 5.1 LLM-based Applications

DPO has emerged as a powerful paradigm for aligning LLMs with human preferences across diverse applications [116, 235, 263, 264]. In code generation, DPO enhances control over code quality by optimizing based on preferences from automated tests [265, 266, 267]. In mathematical reasoning, DPO reduces errors in complex problem-solving by emphasizing step-level preference optimization [100, 101, 129, 268]. Multilingual applications leverage DPO to synchronize cross-lingual preferences, thereby improving translation accuracy and cultural relevance [107, 269]. Recommendation systems utilize DPO to refine personalization by incorporating user preference data to optimize item rankings, thereby enhancing the model ability to distinguish preferred items from less preferred ones [270, 271]. These examples highlight the adaptability of DPO in achieving human-aligned outputs across diverse tasks.

### 5.2 Diffusion Models

In the realm of diffusion models, DPO has been adapted to better align generated content with user expectations [272, 273, 274, 275]. By optimizing preferences over image-text pairs, DPO enhances the semantic accuracy of generated images and mitigates the production of undesirable or biased content. Studies have demonstrated that diffusion models fine-tuned with DPO respond more accurately to complex prompts compared to those trained with traditional techniques. Moreover, the efficiency of DPO allows for the fine-tuning of large-scale models using limited preference data, addressing significant computational challenges in training diffusion models [276, 277, 278]. While scaling DPO for high-resolution and dynamic content generation remains

challenging, its ability to simplify reward modeling makes it a promising method for controlled content creation [279].

### 5.3 Multi-Modal LLMs

For multi-modal LLMs, DPO plays a crucial role in aligning preferences across different data types, thereby improving coherence in tasks such as visual question answering and image captioning [89, 280, 281, 282, 283]. By optimizing alignment between textual responses and visual inputs, DPO reduces hallucinations in multi-modal interactions, ensuring outputs remain faithful to the given context. Although reconciling different types of feedback can be challenging, DPO offers a practical framework for lightweight adaptation, making it well-suited to preference-intensive multi-modal applications [280, 284, 285].

## 6 CHALLENGES AND FUTURE DIRECTIONS

In this section, we discuss the key challenges and future directions in DPO research. We identify several critical issues that need to be addressed to further advance the field. Moreover, we propose several promising research directions that can help overcome these challenges and accelerate the adoption of DPO in the future.

### 6.1 Efficient Preference Optimization

Efficient preference optimization remains a pivotal challenge, as current DPO methods hinge on the availability of high-quality preference data, yet the manual collection of human annotations is both time-consuming and labor-intensive while automatically model-generated datasets often suffer from issues such as limited diversity, inherent biases, and insufficient fidelity to human judgment [121, 122, 128, 129]. Moreover, even though DPO circumvents the intricacies of reward model engineering common in RL, it does not fully leverage the exploratory strengths that RL methods offer, as evidenced by recent advances in reasoning approaches where RL-based training has achieved notable successes [18, 19]. This opens up an avenue for future research to not only enhance data efficiency through advanced learning techniques but also to integrate novel exploration mechanisms [138, 141], potentially through hybrid models that amalgamate the direct preference optimization benefits of DPO with the robust exploratory capabilities characteristic of RL.

### 6.2 Multi-Modal Preference Optimization

Multi-Modal Preference Optimization presents another frontier, given that existing DPO frameworks have primarily targeted text-based modalities while many real-world applications demand the alignment of diverse human preferences across text, images, audio, and even video [280, 284, 285, 286, 287]. In scenarios where cross-modal cues might conflict, such as the need for concise text paired with richly detailed imagery, the challenge lies in constructing a unified preference representation space that can intelligently and automatically recalibrate the priority of different modalities based on the contextual demands of the task at hand [89, 282, 283]. Future directions in this area could involve the development of innovative multi-modal preference encoding architectures,

which are capable of disentangling compound preferences into modality-specific and cross-modal components that align conflicting preferences while also adapting dynamically to changing inputs.

### 6.3 Continuous Preference Optimization

Continuous preference optimization addresses the dynamic nature of human preferences that evolve over time or vary with different phases of a task, a factor that static DPO models often fail to capture [123, 135, 137, 185]. As social norms and individual preferences shift, there is an increasing need for systems that can continuously recalibrate their alignment strategies in real time while simultaneously mitigating the risk of catastrophic forgetting. Future research in this domain may focus on meta-learning approaches that enable models to learn not only from the current state of preferences but also how to efficiently adapt when these preferences change. By integrating online learning frameworks with mechanisms for detecting temporal shifts and contextual variability in user behavior, researchers can pave the way toward systems that remain consistently relevant and effective in the face of evolving societal and individual expectations.

### 6.4 Interpretable Preference Optimization

Interpretable preference optimization is critical for building trust in models that implicitly align human values, as the opaque nature of current DPO complicates the ability to audit and control the alignment process. In practice, human preferences are multi-dimensional [150, 151, 154], encompassing aspects such as factual accuracy, fairness, creativity, and beyond, and there is a pressing need to decompose these complex preferences into interpretable components that can be individually examined and fine-tuned. Future research could leverage advances in explainable techniques to develop models that not only achieve fine-grained alignment across diverse values but also provide transparent insights into how different preference dimensions interact to shape final decisions. This level of interpretability would allow stakeholders to balance competing values more effectively, ensuring that the alignment process remains both accountable and adaptable as societal norms continue to evolve.

## 7 CONCLUSION

In recent years, DPO has emerged as a promising paradigm for aligning LLMs with human preferences by directly optimizing model policies using preference data. Despite its potential, the DPO research landscape remains fragmented, with a lack of systematic organization and comparative analysis. In this survey, we present a comprehensive overview of DPO and introduce a novel taxonomy that categorizes existing works into four key dimensions: data strategy, learning framework, constraint mechanism, and model property. We have also discussed the key benchmarks, evaluation results, and applications of DPO, highlighting the challenges and future directions in this field. By providing a systematic analysis of the existing DPO methods, we aim to facilitate further research and development in this area.

## REFERENCES

- [1] Wayne Xin Zhao et al. A survey of large language models. *arXiv*, 2023.
- [2] Humza Naveed et al. A comprehensive overview of large language models. *arXiv*, 2023.
- [3] Yupeng Chang et al. A survey on evaluation of large language models. *TIIS*, 2024.
- [4] Shervin Minaee et al. Large language models: A survey. *arXiv*, 2024.
- [5] Shukang Yin et al. A survey on multimodal large language models. *arXiv*, 2023.
- [6] Duzhen Zhang et al. Mm-llms: Recent advances in multimodal large language models. *ACL*, 2024.
- [7] Jingyi Zhang et al. Vision-language models for vision tasks: A survey. *TPAMI*, 2024.
- [8] Zhehui Wang et al. Enabling energy-efficient deployment of large language models on memristor crossbar: A synergy of large and small. *TPAMI*, 2024.
- [9] Hongru Wang et al. A survey of the evolution of language model-based dialogue systems. *arXiv*, 2023.
- [10] Zihao Yi et al. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv*, 2024.
- [11] Jiawei Liu et al. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *NeurIPS*, 2023.
- [12] Daya Guo et al. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. *arXiv*, 2024.
- [13] Xue Jiang et al. Self-planning code generation with large language models. *TOSEM*, 2024.
- [14] Dave Van Veen et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 2024.
- [15] Jesuotofunmi A Omiye et al. Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of Internal Medicine*, 2024.
- [16] Karan Singhal et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 2025.
- [17] Fenglin Liu et al. Aligning, autoencoding and prompting large language models for novel disease reporting. *TPAMI*, 2025.
- [18] Aaron Jaech et al. Openai o1 system card. *arXiv*, 2024.
- [19] Daya Guo et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, 2025.
- [20] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 2015.
- [21] Xiaowei Huang et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 2024.
- [22] Yue Zhang et al. Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv*, 2023.
- [23] Isabel O Gallegos et al. Bias and fairness in large language models: A survey. *Computational Linguistics*, 2024.
- [24] Yufei Wang et al. Aligning large language models with human: A survey. *arXiv*, 2023.
- [25] Yang Liu et al. Trustworthy llms: A survey and guideline for evaluating large language models’ alignment. *arXiv*, 2023.
- [26] Tianhao Shen et al. Large language model alignment: A survey. *arXiv*, 2023.
- [27] Hannah Rose Kirk et al. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 2024.
- [28] Usman Anwar et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv*, 2024.
- [29] Bofei Gao et al. Towards a unified view of preference learning for large language models: A survey. *arXiv*, 2024.
- [30] Ruili Jiang et al. A survey on human preference learning for large language models. *arXiv*, 2024.
- [31] Zhichao Wang et al. A comprehensive survey of llm alignment techniques: Rlhf, rlai, ppo, dpo and more. *arXiv*, 2024.
- [32] Genta Indra Winata et al. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *arXiv*, 2024.
- [33] Yue Huang et al. Position: TrustLLM: Trustworthiness in large language models. *ICML*, 2024.
- [34] Paul F Christiano et al. Deep reinforcement learning from human preferences. *NeurIPS*, 2017.
- [35] Long Ouyang et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.
- [36] Nisan Stiennon et al. Learning to summarize with human feedback. *NeurIPS*, 2020.
- [37] Josh Achiam et al. Gpt-4 technical report. *arXiv*, 2023.
- [38] Yuntao Bai et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*, 2022.
- [39] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024.
- [40] Yuchun Miao et al. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *NeurIPS*, 2024.
- [41] Stephen Casper et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv*, 2023.
- [42] Keertana Chidambaram et al. Direct preference optimization with unobserved preference heterogeneity. *arXiv*, 2024.
- [43] Haoxian Chen et al. Mallowspo: Fine-tune your llm with preference dispersions. *arXiv*, 2024.
- [44] Shyam Sundhar Ramesh et al. Group robust preference optimization in reward-free rlhf. *arXiv*, 2024.
- [45] Binwei Yao et al. No preference left behind: Group distributional preference optimization. *ICLR*, 2025.
- [46] Afra Amini et al. Direct preference optimization with an offset. *ACL Findings*, 2024.
- [47] Qi Gou and Cam-Tu Nguyen. Mixed preference optimization: Reinforcement learning with data selection and better reference model. *arXiv*, 2024.
- [48] Shiqi Wang et al. Reward difference optimization for sample reweighting in offline RLHF. *EMNLP Findings*, 2024.
- [49] Junkang Wu et al.  $\alpha$ -dpo: Adaptive reward margin is what direct preference optimization needs. *arXiv*, 2024.
- [50] Hiroki Furuta et al. Geometric-averaged preference optimization for soft preference labels. *NeurIPS*, 2024.
- [51] Junkang Wu et al. Beta-dpo: Direct preference optimization with dynamic beta. *NeurIPS*, 2024.
- [52] Tetsuro Morimura et al. Filtered direct preference optimization. *EMNLP*, 2024.
- [53] Pulkit Pattnaik et al. Enhancing alignment using curriculum learning & ranked preferences. *EMNLP*, 2024.
- [54] Ilgee Hong et al. Adaptive preference scaling for reinforcement learning with human feedback. *NeurIPS*, 2024.
- [55] Dahyun Kim et al. Sdpo: Don’t use your data all at once. *arXiv*, 2024.
- [56] Runsheng Yu et al. Direct alignment of language models via quality-aware self-refinement. *arXiv*, 2024.
- [57] Lou Jieming et al. Gap-aware preference optimization: Enhancing model alignment with perception margin. *OpenReview*, 2024.
- [58] Jingyuan Ma et al. Plug-and-play training framework for preference optimization. *arXiv*, 2024.
- [59] Sayak Ray Chowdhury et al. Provably robust DPO: Aligning language models with noisy feedback. *ICML*, 2024.
- [60] Keyi Kong et al. Perplexity-aware correction for robust alignment with noisy preferences. *NeurIPS*, 2024.
- [61] Xize Liang et al. Ropo: Robust preference optimization for large language models. *arXiv*, 2024.
- [62] Dongyoung Kim et al. Spread preference annotation: Direct preference judgment for efficient LLM alignment. *ICLR*, 2025.
- [63] Lingfan Zhang et al. Combating inherent noise for direct preference optimization. *OpenReview*, 2025.
- [64] Shawn Im and Yixuan Li. Understanding generalization of preference optimization under noisy feedback. *OpenReview*, 2025.
- [65] Yang Gao et al. Impact of preference noise on the alignment performance of generative language models. *COLM*, 2024.
- [66] Junkang Wu et al. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. *ICLR*, 2024.
- [67] Kawin Ethayarajh et al. Model alignment as prospect theoretic optimization. *ICML*, 2024.
- [68] Seungjae Jung et al. Binary classifier optimization for large language model alignment. *arXiv*, 2024.
- [69] Teng Xiao et al. Cal-dpo: Calibrated direct preference optimization for language model alignment. *NeurIPS*, 2024.
- [70] Igor Melnyk et al. Distributional preference alignment of llms via optimal transport. *NeurIPS*, 2024.
- [71] Tianchi Cai et al. Ulma: Unified language model alignment with human demonstration and point-wise preference. *arXiv*, 2023.
- [72] Huayu Chen et al. Noise contrastive alignment of language models with explicit rewards. *NeurIPS*, 2024.
- [73] Yifan Zhang et al. General preference modeling with preference representations for aligning language models. *arXiv*, 2024.
- [74] Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

- [75] Mohammad Gheshlaghi Azar et al. A general theoretical paradigm to understand learning from human preferences. *AISTATS*, 2024.
- [76] Jinghong Chen et al. On extending direct preference optimization to accommodate ties. *arXiv*, 2024.
- [77] Yuxin Jiang et al. Bridging and modeling correlations in pairwise data for direct preference optimization. *arXiv*, 2024.
- [78] Xinghua Zhang et al. Iopo: Empowering llms with complex instruction following via input-output preference optimization. *arXiv*, 2024.
- [79] Abbas Abdolmaleki et al. Preference optimization as probabilistic inference. *ICLR*, 2024.
- [80] Yueqin Yin et al. Self-augmented preference optimization: Off-policy paradigms for language model alignment. *arXiv*, 2024.
- [81] Shitong Duan et al. Negating negatives: Alignment with human negative samples via distributional dispreference optimization. *arXiv*, 2024.
- [82] Ruiqi Zhang et al. Negative preference optimization: From catastrophic collapse to effective unlearning. *COLM*, 2024.
- [83] Chongyu Fan et al. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv*, 2024.
- [84] Yifan Zhong et al. Panacea: Pareto alignment via preference adaptation for llms. *NeurIPS*, 2024.
- [85] Tianqi Liu et al. Lipo: Listwise preference optimization through learning-to-rank, 2024. *arXiv*, 2024.
- [86] Mingye Zhu et al. LIRE: listwise reward enhancement for preference alignment. *ACL*, 2024.
- [87] Yang Zhao et al. Ordinal preference optimization: Aligning human preferences via ndcg. *arXiv*, 2024.
- [88] Jiacong Zhou et al. Optimizing preference alignment with differentiable ndcg ranking. *arXiv*, 2024.
- [89] Fei Wang et al. mDPO: Conditional preference optimization for multimodal large language models. *EMNLP*, 2024.
- [90] Yueqin Yin et al. Relative preference optimization: Enhancing llm alignment through contrasting responses across identical and diverse prompts. *arXiv*, 2024.
- [91] Yuxiang Guo et al. Todo: Enhancing llm alignment with ternary preferences. *ICLR*, 2024.
- [92] Rafael Rafailov et al. From r to q\*: Your language model is secretly a q-function. *COLM*, 2024.
- [93] Yongcheng Zeng et al. Token-level direct preference optimization. *ICML*, 2024.
- [94] Aiwei Liu et al. Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights. *ICLR*, 2024.
- [95] Fenia Christopoulou et al. Sparsepo: Controlling preference alignment of llms via sparse token masks. *arXiv*, 2024.
- [96] Han Zhong et al. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv*, 2024.
- [97] Kailai Yang et al. Selective preference optimization via token-level reward function estimation. *arXiv*, 2024.
- [98] Qi Zhao et al. EPO: hierarchical LLM agents with environment preference optimization. *EMNLP*, 2024.
- [99] Ruichen Shao et al. Earlier tokens contribute more: Learning direct preference optimization from temporal decay perspective. *ICLR*, 2025.
- [100] Xin Lai et al. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv*, 2024.
- [101] Zimu Lu et al. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning. *arXiv*, 2024.
- [102] Xuan Zhang et al. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *NeurIPS*, 2024.
- [103] Yuxi Xie et al. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv*, 2024.
- [104] Weibin Liao et al. Tpo: Aligning large language models with multi-branch & multi-step preference trees. *arXiv*, 2024.
- [105] Hoang Anh Just et al. Data-centric human preference optimization with rationales. *arXiv*, 2024.
- [106] Jiakai Liu et al. Improving multi-step reasoning abilities of large language models with direct advantage policy optimization. *arXiv*, 2024.
- [107] Shuaijie She et al. MAPO: advancing multilingual reasoning through multilingual-alignment-as-preference optimization. *ACL*, 2024.
- [108] Lifan Yuan et al. Advancing llm reasoning generalists with preference trees. *arXiv*, 2024.
- [109] Richard Yuanzhe Pang et al. Iterative reasoning preference optimization. *NeurIPS*, 2024.
- [110] Chao-Wei Huang and Yun-Nung Chen. Factalign: Long-form factuality alignment of large language models. *arXiv*, 2024.
- [111] Wei Xiong et al. Building math agents with multi-turn iterative preference learning. *ICLR*, 2025.
- [112] Yifan Song et al. Trial and error: Exploration-based trajectory optimization for llm agents. *ACL*, 2024.
- [113] Aobo Kong et al. Sdpo: Segment-level direct preference optimization for social agents. *arXiv*, 2025.
- [114] Pranav Putta et al. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv*, 2024.
- [115] Wentao Shi et al. Direct multi-turn preference optimization for language agents. *EMNLP*, 2024.
- [116] Haoran Xu et al. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *ICML*, 2024.
- [117] Jiwoo Hong et al. ORPO: Monolithic preference optimization without reference model. *EMNLP*, 2024.
- [118] Shiva Kumar Pentiyala et al. Paft: A parallel training paradigm for effective llm fine-tuning. *arXiv*, 2024.
- [119] Songyang Gao et al. Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback. *ICML*, 2024.
- [120] Feifan Song et al. Icdpo: Effectively borrowing alignment capability of others via in-context direct preference optimization. *arXiv*, 2024.
- [121] Shangmin Guo et al. Direct language model alignment from online ai feedback. *arXiv*, 2024.
- [122] Biqing Qi et al. Online dpo: Online direct preference optimization with fast-slow chasing. *arXiv*, 2024.
- [123] Weizhe Yuan et al. Self-rewarding language models. *ICML*, 2024.
- [124] Wenda Xu et al. BPO: Staying close to the behavior LLM creates better online LLM alignment. *EMNLP*, 2024.
- [125] Saeed Khaki et al. RS-DPO: A hybrid rejection sampling and direct preference optimization method for alignment of large language models. *NAACL*, 2024.
- [126] Tianqi Liu et al. Statistical rejection sampling improves preference optimization. *ICLR*, 2024.
- [127] Ruizhe Shi et al. The crucial role of samplers in online direct preference optimization. *ICLR*, 2025.
- [128] Lichang Chen et al. Optune: Efficient online preference tuning. *arXiv*, 2024.
- [129] Tianduo Wang et al. Self-training with direct preference optimization improves chain-of-thought reasoning. *ACL*, 2024.
- [130] Jiafan He et al. Accelerated preference optimization for large language model alignment. *arXiv*, 2024.
- [131] Wei Xiong et al. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. *ICML*, 2024.
- [132] Yixin Liu et al. Comal: A convergent meta-algorithm for aligning llms with general preferences. *arXiv*, 2024.
- [133] Jing Xu et al. Some things are more cringe than others: Iterative preference optimization with the pairwise cringe loss. *arXiv*, 2024.
- [134] Jongwoo Ko et al. Sera: Self-reviewing and alignment of large language models using implicit reward margins. *ICLR*, 2025.
- [135] Zhaoyang Wang et al. Cream: Consistency regularized self-rewarding language models. *ICLR*, 2025.
- [136] Prasann Singhal et al. D2PO: Discriminator-guided DPO with response evaluation models. *COLM*, 2024.
- [137] Aiwei Liu et al. Direct large language model alignment through self-rewarding contrastive prompt distillation. *ACL*, 2024.
- [138] Tengyang Xie et al. Exploratory preference optimization: Provably sample-efficient exploration in rlhf with general function approximation. *ICLR*, 2025.
- [139] Shengao Zhang et al. Self-exploring language models: Active preference elicitation for online alignment. *arXiv*, 2024.
- [140] Shicong Cen et al. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *ICLR*, 2025.
- [141] Chenjia Bai et al. Online preference alignment for language models via count-based exploration. *ICLR*, 2025.
- [142] Yuda Song et al. The importance of online data: Understanding preference fine-tuning via coverage. *NeurIPS*, 2024.
- [143] Yaojie Shen et al. Aipo: Improving training objective for iterative preference optimization. *arXiv*, 2024.
- [144] Yunhao Tang et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv*, 2024.
- [145] Shusheng Xu et al. Is DPO superior to PPO for LLM alignment? A comprehensive study. *ICML*, 2024.
- [146] William Muldrew et al. Active preference learning for large

- language models. *ICML*, 2024.
- [147] Seola Choi et al. Active preference optimization via maximizing learning capacity. *OpenReview*, 2024.
- [148] Kaixuan Ji et al. Reinforcement learning from human feedback with active queries. *arXiv*, 2024.
- [149] Nirjhar Das et al. Active preference optimization for sample efficient rlhf. *arXiv*, 2024.
- [150] Zhanhui Zhou et al. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. *ACL Findings*, 2024.
- [151] Xingzhou Lou et al. Spo: Multi-dimensional preference sequential alignment with implicit reward modeling. *arXiv*, 2024.
- [152] Yu Zhang et al. MOSLIM: Align with diverse preferences in prompts through reward classification. *OpenReview*, 2025.
- [153] Anirudhan Badrinath et al. Hybrid preference optimization: Augmenting direct preference optimization with auxiliary objectives. *arXiv*, 2024.
- [154] Yiju Guo et al. Controllable preference optimization: Toward controllable multi-objective alignment. *EMNLP*, 2024.
- [155] Abhijnan Nath et al. Simultaneous reward distillation and preference learning: Get you a language model who can do both. *arXiv*, 2024.
- [156] Zixiang Chen et al. Self-play fine-tuning converts weak language models to strong language models. *ICML*, 2024.
- [157] Yue Wu et al. Self-play preference optimization for language model alignment. *ICLR*, 2025.
- [158] Gokul Swamy et al. A minimaximalist approach to reinforcement learning from human feedback. *ICML*, 2024.
- [159] Lin Gui et al. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *NeurIPS*, 2024.
- [160] Remi Munos et al. Nash learning from human feedback. *ICML*, 2024.
- [161] Corby Rosset et al. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv*, 2024.
- [162] Daniele Calandriello et al. Human alignment of large language models through online preference optimisation. *ICML*, 2024.
- [163] Eugene Choi et al. Self-improving robust preference optimization. *ICLR*, 2025.
- [164] Haoyan Yang et al. Dynamic noise preference optimization for llm self-improvement via synthetic data. *arXiv*, 2025.
- [165] Alexey Gorbatoevski et al. Learn your reference model for real good alignment. *arXiv*, 2024.
- [166] Yu Meng et al. Simpo: Simple preference optimization with a reference-free reward. *NeurIPS*, 2024.
- [167] Teng Xiao et al. SimPER: A minimalist approach to preference alignment without hyperparameters. *ICLR*, 2025.
- [168] Yixin Liu et al. Understanding reference policies in direct preference optimization. *arXiv*, 2024.
- [169] Chaoqi Wang et al. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *ICLR*, 2023.
- [170] Stewart Slocum et al. Diverse preference learning for capabilities and alignment. *ICLR*, 2025.
- [171] Amitava Das et al. Dpo kernels: A semantically-aware, kernel-enhanced, and divergence-rich paradigm for direct preference optimization. *arXiv*, 2025.
- [172] Mingye Zhu et al. FlipGuard: Defending preference alignment against update regression with constrained optimization. *EMNLP*, 2024.
- [173] Qingyu Yin et al. Direct preference optimization using sparse feature-level constraints. *arXiv*, 2024.
- [174] Yunhao Tang et al. Generalized preference optimization: A unified approach to offline alignment. *ICML*, 2024.
- [175] Haozhe Ji et al. Towards efficient exact optimization of language model alignment. *ICML*, 2024.
- [176] Arsalan Sharifnassab et al. Soft preference optimization: Aligning language models to expert distributions. *arXiv*, 2024.
- [177] Janghwan Lee et al. Improving conversational abilities of quantized large language models via direct preference alignment. *ACL*, 2024.
- [178] Audrey Huang et al. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv*, 2025.
- [179] Geon-Hyeong Kim et al. SafeDPO: A simple approach to direct preference optimization with enhanced safety. *OpenReview*, 2025.
- [180] Akifumi Wachi et al. Stepwise alignment for constrained language model policy optimization. *NeurIPS*, 2024.
- [181] Zixuan Liu et al. Enhancing llm safety via constrained direct preference optimization. *arXiv*, 2024.
- [182] San Kim and Gary Geunbae Lee. Adversarial dpo: Harnessing harmful data for reducing toxicity with minimal impact on coherence and evasiveness in dialogue agents. *arXiv*, 2024.
- [183] Andrew Lee et al. A mechanistic understanding of alignment algorithms: a case study on dpo and toxicity. *ICML*, 2024.
- [184] Yiming Zhang et al. Backtracking improves generation safety. *ICLR*, 2025.
- [185] Seongho Son et al. Right now, wrong then: Non-stationary direct preference optimization under preference drift. *arXiv*, 2024.
- [186] Eugene Choi et al. Self-improving robust preference optimization. *ICLR*, 2025.
- [187] Adam Fisch et al. Robust preference optimization through reward model distillation. *arXiv*, 2024.
- [188] Yong Lin et al. On the limited generalization capability of the implicit reward model induced by direct preference optimization. *EMNLP Findings*, 2024.
- [189] Fahim Tajwar et al. Preference fine-tuning of llms should leverage suboptimal, on-policy data. *ICML*, 2024.
- [190] Hongyi Yuan et al. Rrhf: Rank responses to align language models with human feedback. *NeurIPS*, 2023.
- [191] Ryan Park et al. Disentangling length from quality in direct preference optimization. *ACL Findings*, 2024.
- [192] Junru Lu et al. Eliminating biased length reliance of direct preference optimization via down-sampled KL divergence. *EMNLP*, 2024.
- [193] Weizhe Yuan et al. Following length constraints in instructions. *arXiv*, 2024.
- [194] Kian Ahrabian et al. The hitchhiker’s guide to human alignment with\* po. *arXiv*, 2024.
- [195] Wei Liu et al. Length desensitization in directed preference optimization. *arXiv*, 2024.
- [196] Guanzheng Chen et al. LongPO: Long context self-evolution of large language models through short-to-long preference optimization. *ICLR*, 2025.
- [197] Prasann Singhal et al. A long way to go: Investigating length correlations in RLHF. *COLM*, 2024.
- [198] Kyle Richardson et al. Understanding the logic of direct preference alignment through logic. *arXiv*, 2024.
- [199] Karel D’Oosterlinck et al. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment. *arXiv*, 2024.
- [200] Arka Pal et al. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv*, 2024.
- [201] Yuzi Yan et al. 3d-properties: Identifying challenges in DPO and charting a path forward. *ICLR*, 2025.
- [202] Duanyu Feng et al. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv*, 2024.
- [203] Hui Yuan et al. A common pitfall of margin-based language model alignment: Gradient entanglement. *ICLR*, 2025.
- [204] Noam Razin et al. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv*, 2024.
- [205] Zhengyan Shi et al. Understanding likelihood over-optimisation in direct alignment algorithms. *arXiv*, 2024.
- [206] Yong Lin et al. Mitigating the alignment tax of RLHF. *EMNLP*, 2024.
- [207] Megh Thakkar et al. A deep dive into the trade-offs of parameter-efficient preference alignment techniques. *ACL*, 2024.
- [208] Keming Lu et al. Online merging optimizers for boosting rewards and mitigating tax in alignment. *arXiv*, 2024.
- [209] Angelica Chen et al. Preference learning algorithms do not learn preference rankings. *NeurIPS*, 2024.
- [210] Wenyi Xiao et al. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications. *arXiv*, 2024.
- [211] Pierre Harvey Richemond et al. Offline regularised reinforcement learning for large language models alignment. *arXiv*, 2024.
- [212] Christian Wirth et al. A survey of preference-based reinforcement learning methods. *JMLR*, 2017.
- [213] Jiaming Ji et al. Ai alignment: A comprehensive survey. *arXiv*, 2023.
- [214] Xinpeng Wang et al. On the essence and prospect: An investigation of alignment approaches for big models. *IJCAI*, 2024.
- [215] Hannah Rose Kirk et al. The past, present and better future of feedback learning in large language models for subjective human preferences and values. *EMNLP*, 2023.
- [216] Patrick Fernandes et al. Bridging the gap: A survey on integrating

- (human) feedback for natural language generation. *TACL*, 2023.
- [217] Timo Kaufmann et al. A survey of reinforcement learning from human feedback. *arXiv*, 2023.
- [218] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.
- [219] John Schulman et al. Proximal policy optimization algorithms. *arXiv*, 2017.
- [220] Arash Ahmadian et al. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *ACL*, 2024.
- [221] Ziniu Li et al. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *ICML*, 2024.
- [222] Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv*, 2024.
- [223] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv*, 2025.
- [224] Chris Lu et al. Discovering preference optimization algorithms with and for large language models. *NeurIPS*, 2024.
- [225] Hanyang Zhao et al. RainbowPO: A unified framework for combining improvements in preference optimization. *ICLR*, 2025.
- [226] Hamish Ivison et al. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *NeurIPS*, 2024.
- [227] Amir Saeidi et al. Insights into alignment: Evaluating dpo and its variants across multiple tasks. *arXiv*, 2024.
- [228] Andi Nika et al. Reward model learning vs. direct policy optimization: a comparative analysis of learning from human preferences. *ICML*, 2024.
- [229] Ziniu Li et al. When is rl better than dpo in rlhf? a representation and optimization perspective. *ICLR Tiny Papers*, 2024.
- [230] Yao Zhao et al. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv*, 2023.
- [231] Feifan Song et al. Preference ranking optimization for human alignment. *AAAI*, 2024.
- [232] Chaoqi Wang et al. Preference optimization with multi-sample comparisons. *arXiv*, 2024.
- [233] Ziniu Li et al. Policy optimization in rlhf: The impact of out-of-preference data. *arXiv*, 2023.
- [234] Lei Li et al. Improving reasoning ability of large language models via iterative uncertainty-based preference optimization. *OpenReview*, 2025.
- [235] Abhimanyu Dubey et al. The llama 3 herd of models. *arXiv*, 2024.
- [236] Lily H Zhang and Rajesh Ranganath. Win rate is all that can matter from preference data alone. *OpenReview*, 2025.
- [237] Ganqu Cui et al. Ultrafeedback: Boosting language models with high-quality feedback. *ICML*, 2023.
- [238] Jiaming Ji et al. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv*, 2024.
- [239] Zhilin Wang et al. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv*, 2023.
- [240] Hunter Lightman et al. Let’s verify step by step. *ICLR*, 2023.
- [241] Kawin Ethayarajh et al. Understanding dataset difficulty with v-usable information. *ICML*, 2022.
- [242] Banghua Zhu et al. Starling-7b: Improving llm helpfulness & harmlessness with rlhf, 2023.
- [243] Wing Lian et al. Openorca: An open dataset of gpt augmented flan reasoning traces, 2023.
- [244] Luigi Daniele and Suphadeeprasit. Amplify-instruct: Synthetically generated diverse multi-turn conversations for efficient llm training., 2023.
- [245] Jiaming Ji et al. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *NeurIPS*, 2023.
- [246] Andrew Maas et al. Learning word vectors for sentiment analysis. *ACL*, 2011.
- [247] Michael Völske et al. Tl; dr: Mining reddit to learn automatic summarization. *EMNLP Workshop*, 2017.
- [248] Deep Ganguli et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv*, 2022.
- [249] Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv*, 2021.
- [250] Yann Dubois et al. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv*, 2024.
- [251] Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023.
- [252] Andy Zou et al. Universal and transferable adversarial attacks on aligned language models. *arXiv*, 2023.
- [253] Tianle Li et al. From live data to high-quality benchmarks: The arena-hard pipeline. 2024.
- [254] Stephanie Lin et al. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv*, 2021.
- [255] Jeffrey Zhou et al. Instruction-following evaluation for large language models. *arXiv*, 2023.
- [256] Mirac Suzgun et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv*, 2022.
- [257] Dan Hendrycks et al. Measuring mathematical problem solving with the math dataset. *arXiv*, 2021.
- [258] David Rein et al. Gpqa: A graduate-level google-proof q&a benchmark. *COLM*, 2024.
- [259] Zayne Sprague et al. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv*, 2023.
- [260] Yubo Wang et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *NeurIPS*, 2024.
- [261] Fengqing Jiang et al. Identifying and mitigating vulnerabilities in llm-integrated applications. *arXiv*, 2023.
- [262] Ning Ding et al. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv*, 2023.
- [263] Qiyu Wu et al. Word alignment as preference for machine translation. *EMNLP*, 2024.
- [264] Yinghao Hu et al. Fine-tuning large language models for improving factuality in legal question answering. *COLING*, 2025.
- [265] Leonidas Gee et al. Code-optimize: Self-generated preference data for correctness and efficiency. *arXiv*, 2024.
- [266] Yibo Miao et al. Aligning codellms with direct preference optimization. *arXiv*, 2024.
- [267] Kechi Zhang et al. Codedpo: Aligning code models with self generated and verified source code. *arXiv*, 2024.
- [268] Guoxin Chen et al. Step-level value preference optimization for mathematical reasoning. *EMNLP*, 2024.
- [269] Wen Lai et al. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. *ACL Findings*, 2024.
- [270] Yuxin Chen et al. On softmax direct preference optimization for recommendation. *NeurIPS*, 2024.
- [271] Zhuoxi Bai et al. Finetuning large language model for personalized ranking. *arXiv*, 2024.
- [272] Yi Gu et al. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv*, 2024.
- [273] Shivanshu Shekhar et al. See-dpo: Self entropy enhanced direct preference optimization. *arXiv*, 2024.
- [274] Shufan Li et al. Aligning diffusion models by optimizing human utility. *NeurIPS*, 2024.
- [275] Navonil Majumder et al. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *ACM MM*, 2024.
- [276] Bram Wallace et al. Diffusion model alignment using direct preference optimization. *CVPR*, 2024.
- [277] Shentao Yang et al. A dense reward view on aligning text-to-image diffusion with preference. *ICML*, 2024.
- [278] Kai Yang et al. Using human feedback to fine-tune diffusion models without any reward model. *CVPR*, 2024.
- [279] Buhua Liu et al. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv*, 2024.
- [280] Shengzhi Li et al. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. *ACL*, 2024.
- [281] Ziqi Liang et al. AlignCap: Aligning speech emotion captioning to human preferences. *EMNLP*, 2024.
- [282] Elmira Amirloo et al. Understanding alignment in multimodal llms: A comprehensive study. *arXiv*, 2024.
- [283] Jinlan Fu et al. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *arXiv*, 2025.
- [284] Ruohong Zhang et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv*, 2024.
- [285] Yuxi Xie et al. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *EMNLP Findings*, 2024.
- [286] Peng Xu et al. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *TPAMI*, 2025.
- [287] Zhongzhan Huang et al. A causality-aware paradigm for evaluating creativity of multimodal large language models. *TPAMI*, 2025.