Novelty Assessment Report

Paper: SafeDPO: A Simple Approach to Direct Preference Optimization with Enhanced Safety
PDF URL: https://openreview.net/pdf?id=PJdw4VBsXD
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

As Large Language Models (LLMs) are increasingly deployed in real-world applications, balancing both helpfulness and safety has become a central challenge. A natural approach is to incorporate safety constraints into Reinforcement Learning from Human Feedback (RLHF), where recent studies have shown promising progress. However, these methods often rely on auxiliary networks or multi-stage pipelines, thereby increasing complexity. In this work, we revisit the safety alignment objective itself and demonstrate that it admits a closed-form solution, yielding a theoretically grounded and provably equivalent reformulation that enables a direct and tractable optimization procedure. Building on this insight, we propose SafeDPO, a lightweight method derived from this formulation, which preserves the optimality of the underlying safety-constrained objective while requiring only one additional hyperparameter and minimal modifications to existing preference-based training methods. At the same time, it eliminates the need for reward models, cost models, and online sampling. Despite its simplicity, SafeDPO achieves comparable or superior results to state-of-the-art safety alignment methods in both theoretical soundness and empirical performance. Experiments on the PKU-SafeRLHF-30K benchmark show that SafeDPO consistently improves safety while maintaining competitive helpfulness. Ablation studies further show that the additional hyperparameter provides a flexible mechanism to enhance safety without altering the theoretical optimum, and confirm that SafeDPO scales reliably to LLMs with up to 13B parameters. Overall, our results highlight that a simple, theory-driven objective can provide a lightweight yet effective solution for safety alignment in practice.

Core Task Landscape

This paper addresses: **Safety Alignment in Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Theoretical Foundations and Limitations**
- **Safety Alignment Techniques**
- **Safety Evaluation and Benchmarking**
- **Robustness and Defense Against Safety Attacks**
- **Safety Risks and Vulnerability Analysis**
- **Alignment Beyond Safety**
- **Domain-Specific Applications**

Complete Taxonomy Tree

- Safety Alignment in Large Language Models Survey Taxonomy
- Theoretical Foundations and Limitations
  - Fundamental Alignment Constraints (3 papers)
  - [1] Fundamental limitations of alignment in large language models (Wolf, 2023) View paper
  - [30] Fundamental Safety-Capability Trade-offs in Fine-tuning Large Language Models (Chen, 2025) View paper
  - [49] Safety Alignment Depth in Large Language Models: A Markov Chain Perspective (Yu, 2025) View paper
  - Mechanistic Interpretability of Safety (2 papers)
  - [2] Finding safety neurons in large language models (Jianhui Chen, 2024) View paper
  - [27] Unveiling the Basin-Like Loss Landscape in Large Language Models (Chen, 2025) View paper
- Safety Alignment Techniques
  - Preference-Based Alignment Methods
  - Direct Preference Optimization Variants ★ (2 papers)
    - [0] SafeDPO: A Simple Approach to Direct Preference Optimization with Enhanced Safety (Anon et al., 2026) View paper
    - [29] Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study (Xu Shusheng, 2024) View paper
  - RLHF and Reward-Based Approaches (2 papers)
    - [42] Equilibrate RLHF: Towards Balancing Helpfulness-Safety Trade-off in Large Language Models (Tan Yingshui, 2025) View paper
    - [43] Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge (Tianhao Wu, 2024) View paper
  - Human Preference Datasets for Safety (2 papers)
    - [20] BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset (Ji, 2023) View paper
    - [40] Pku-saferlhf: Towards multi-level safety alignment for llms with human preference (Jiaming Ji, 2025) View paper
  - Fine-Tuning Stage Safety Methods
  - Safe Data Filtering and Curation (2 papers)
    - [3] Safety-aware fine-tuning of large language models (Choi, 2024) View paper

        - [11] Safety fine-tuning at (almost) no cost: A baseline for vision large language models (Zong, 2024) View paper
    - Post-Fine-Tuning Safety Recovery (4 papers)
        - [5] Safe and effective post-fine-tuning alignment in large language models (Minrui Jiang, 2025) View paper
        - [18] Towards comprehensive and efficient post safety alignment of large language models via safety patching (Weixiang Zhao, 2024) View paper
        - [28] Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic (Rishabh Bhardwaj, 2024) View paper
        - [32] EnchTable: Unified Safety Alignment Transfer in Fine-tuned Large Language Models (Wu Jialin, 2025) View paper
    - Gradient-Level Safety Interventions (2 papers)
        - [33] Gradient surgery for safe llm fine-tuning (Yi Biao, 2025) View paper
        - [34] Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack (Sihao Hu, 2024) View paper
    - Parameter-Efficient Safe Fine-Tuning (1 papers)
        - [23] Safe LoRA: the Silver Lining of Reducing Safety Risks when Fine-tuning Large Language Models (Pin-Yu Chen, 2024) View paper
    - Inference-Time Safety Methods (2 papers)
    - [25] Safety Alignment of Large Language Models via Contrasting Safe and Harmful Distributions (Xiaoyun Zhang, 2024) View paper
    - [46] Tuning Language Models by Proxy (Liu, 2024) View paper
- Safety Evaluation and Benchmarking
    - General Safety Evaluation Frameworks (4 papers)
    - [6] Evaluating alignment in large language models: a review of methodologies (Uma E. Sarkar, 2025) View paper
    - [14] Safety Misalignment Against Large Language Models (Gong Yi-chen, 2025) View paper
    - [44] Trustworthy llms: a survey and guideline for evaluating large language models' alignment (Liu Yang, 2023) View paper
    - [48] A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More (Wang Zhichao, 2024) View paper
    - Multimodal Safety Evaluation (4 papers)
    - [7] Mm-safetybench: A benchmark for safety evaluation of multimodal large language models (Xin Liu, 2024) View paper
    - [8] Safebench: A safety evaluation framework for multimodal large language models (Liu, 2026) View paper
    - [10] Safety of multimodal large language models on images and texts (Liu Xin, 2024) View paper
    - [37] From Evaluation to Defense: Advancing Safety in Video Large Language Models (Sun Yi-wei, 2025) View paper
    - Domain-Specific Safety Evaluation (3 papers)
    - [4] Safelawbench: Towards safe alignment of large language models (Chuxue Cao, 2025) View paper
    - [17] Medsafetybench: Evaluating and improving the medical safety of large language models (Chirag Agarwal, 2024) View paper
    - [21] Defining and evaluating physical safety for large language models (Tang, 2024) View paper
- Robustness and Defense Against Safety Attacks
    - Jailbreak Attack Analysis and Defense (3 papers)
    - [12] Multilingual jailbreak challenges in large language models (Deng Yue, 2023) View paper
    - [24] Shaping the safety boundaries: Understanding and defending against jailbreaks in large language models (Lang Gao, 2025) View paper
    - [39] Probing the safety response boundary of large language models via unsafe decoding path generation (Wang Haoyu, 2024) View paper
    - Defense Against Harmful Fine-Tuning (3 papers)
    - [22] Emerging safety attack and defense in federated instruction tuning of large language models (Ye Rui, 2024) View paper
    - [38] Targeted Vaccine: Safety Alignment for Large Language Models Against Harmful Fine-Tuning via Layer-Wise Perturbation (Guozhi Liu, 2024) View paper
    - [45] Antidote: Post-fine-tuning Safety Alignment for Large Language Models against Harmful Fine-tuning (Huang Tiansheng, 2024) View paper
    - Adversarial Robustness in Multimodal Models (1 papers)
    - [35] Robust-LLaVA: On the Effectiveness of Large-Scale Robust Image Encoders for Multi-modal Large Language Models (Malik, 2025) View paper
- Safety Risks and Vulnerability Analysis (6 papers)
    - [9] Ai safety in generative ai large language models: A survey (Li Yun, 2024) View paper
    - [13] A survey of safety and trustworthiness of large language models through the lens of verification and validation (Xiaowei Huang, 2024) View paper
    - [15] Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions (Bianchi, 2023) View paper
    - [16] Navigating the safety landscape: Measuring risks in finetuning large language models (Duen Chau, 2024) View paper
    - [26] Fine-tuning aligned language models compromises safety, even when users do not intend to! (Qi, 2023) View paper
    - [50] On the safety of open-sourced large language models: Does alignment really prevent them from being misused? (Zhang Hangfan, 2023) View paper
- Alignment Beyond Safety (2 papers)
    - [19] Moral alignment for llm agents (Tennant, 2024) View paper
    - [31] Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges (Lu Haoran, 2025) View paper
- Domain-Specific Applications (3 papers)
    - [36] Olmes: A standard for language model evaluations (Yuling Gu, 2025) View paper
    - [41] Towards safe large language models for medicine (T Han, 2024) View paper
    - [47] Trafficsafetygpt: Tuning a pre-trained large language model to a domain-specific expert in transportation safety (Zheng, 2023) View paper

Narrative

Core task: safety alignment in large language models. The field has evolved into a rich ecosystem organized around several major branches. Theoretical Foundations and Limitations explores fundamental questions about what alignment can and cannot achieve, as exemplified by works like Fundamental Limitations Alignment[1]. Safety Alignment Techniques encompasses the methodological core, including preference-based methods such as direct preference optimization variants and reinforcement learning from human feedback approaches like PKU-SafeRLHF[40]. Safety Evaluation and Benchmarking provides the measurement infrastructure through datasets like BeaverTails[20] and domain-specific benchmarks such as MedSafetyBench[17] and SafeLawBench[4]. Robustness and Defense Against

Safety Attacks addresses adversarial concerns including multilingual jailbreaks, while Safety Risks and Vulnerability Analysis investigates how fine-tuning can compromise safety as shown in Fine-Tuning Compromises Safety[26]. Alignment Beyond Safety broadens the scope to moral alignment and other objectives, and Domain-Specific Applications tailors methods to specialized contexts like medicine and traffic safety.

A central tension runs through many branches: the trade-off between model capability and safety, explored in Safety-Capability Trade-offs[30], and the challenge of maintaining alignment during post-training adaptation as studied in Post-Fine-Tuning Alignment[5]. Within preference-based alignment methods, researchers debate the relative merits of different optimization strategies, with some work suggesting DPO Superior PPO[29] in certain settings. SafeDPO[0] sits squarely in this active subfield of direct preference optimization variants, proposing refinements to balance safety objectives more effectively during alignment. Its emphasis contrasts with broader approaches like Safety-Aware Fine-Tuning[3], which addresses safety degradation across diverse fine-tuning scenarios, and complements neighboring work on optimizing the preference learning process itself. The landscape reveals ongoing efforts to make alignment both more robust and more practical across deployment contexts.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study

Authors: Xu Shusheng, Fu Wei, Shusheng Xu, Gao, Jiaxuan, et al. (20 authors total) | **Year/Venue**: 2024 • International Conference on Machine Learning | **URL**: View paper

Abstract

Reinforcement Learning from Human Feedback (RLHF) is currently the most widely used method to align large language models (LLMs) with human preferences. Existing RLHF methods can be roughly categorized as either reward-based or reward-free. Novel applications such as ChatGPT and Claude leverage reward-based methods that first learn a reward model and apply actor-critic algorithms, such as Proximal Policy Optimization (PPO). However, in academic benchmarks, state-of-the-art results are often achi...

Relationship Analysis

Both papers belong to the Direct Preference Optimization Variants category, focusing on reward-free methods for safety alignment in LLMs. While the original paper (SafeDPO) proposes a theoretically grounded extension of DPO that incorporates safety constraints through a closed-form solution and safety-aware preference transformation, the candidate paper conducts a comprehensive comparative study between DPO and PPO methods, examining their algorithmic properties and empirical performance across various benchmarks. The key difference is that SafeDPO extends DPO specifically for safety alignment with theoretical guarantees, whereas the candidate paper evaluates and compares existing DPO and PPO approaches without proposing a new safety-specific method.

Contributions Analysis

Overall novelty summary. The paper proposes SafeDPO, a method for safety-constrained alignment in large language models through a closed-form reformulation of the safety objective. It resides in the 'Direct Preference Optimization Variants' leaf of the taxonomy, which contains only two papers total (including this one). This places the work in a relatively sparse but active research direction within preference-based alignment methods. The sibling paper in this leaf focuses on general DPO improvements, while SafeDPO specifically targets safety-constrained formulations, suggesting a specialized niche within an emerging subfield.

The taxonomy reveals that SafeDPO's parent branch, 'Preference-Based Alignment Methods', contains three distinct approaches: DPO variants, RLHF/reward-based methods, and human preference datasets. Neighboring leaves include RLHF approaches like PKU-SafeRLHF and datasets like BeaverTails that separate helpfulness from harmlessness. The scope note for DPO variants explicitly excludes reward-based methods using critic networks, positioning SafeDPO as part of the reward-free optimization paradigm. Adjacent branches address fine-tuning stage safety and inference-time methods, indicating that SafeDPO operates at the initial alignment stage rather than post-deployment adaptation.

Among the three identified contributions, the closed-form reformulation examined nine candidates with zero refutable prior work, while the SafeDPO algorithm examined ten candidates, also with zero refutations. The safety-aware preference transformation was not examined against any candidates. The limited search scope (19 total candidates examined across all contributions) suggests these findings reflect top-K semantic matches rather than exhaustive coverage. Given the sparse population of the DPO variants leaf and the absence of clear refutations among examined candidates, the core algorithmic contributions appear relatively novel within the constrained search space, though the theoretical reformulation's novelty depends on how it relates to broader optimization literature not captured here.

Based on the limited literature search covering 19 candidates, SafeDPO appears to occupy a distinct position within the emerging DPO-based safety alignment space. The analysis captures semantic neighbors and direct citations but does not exhaustively cover all theoretical optimization work or parallel developments in constrained preference learning. The sparse taxonomy leaf and zero refutations among examined candidates suggest potential novelty, though definitive assessment would require broader coverage of optimization theory and concurrent safety-constrained methods.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Closed-form reformulation of safety alignment objective

Description: The authors derive a tractable, closed-form formulation of the constrained safety alignment problem by reformulating it as an unconstrained optimization problem with a modified reward function. This reformulation eliminates the need for surrogate relaxations or auxiliary models while preserving optimality.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Joint verification and refinement of language models for safety-constrained planning

URL: View paper

Brief Assessment

Verification Refinement Planning[64] focuses on formal verification of robot programs against safety specifications using automaton-based representations and model checking. It does not address closed-form solutions for constrained safety alignment optimization in language models or reformulation of RLHF objectives.

---

2. Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback

URL: View paper

Brief Assessment

Linear Alignment[62] focuses on a closed-form solution for general preference alignment without training or feedback, using linear approximation of policy optimization. The ORIGINAL paper addresses constrained safety alignment with hard safety constraints, reformulating it as an unconstrained problem with modified rewards. These are distinct problem formulations and objectives.

---

3. Alignment of large language models with constrained learning

URL: View paper

Brief Assessment

Constrained Learning[63] focuses on computing optimal dual variables and characterizing primal-dual gaps in constrained alignment, rather than deriving closed-form reformulations that eliminate surrogate relaxations. The candidate employs Lagrangian duality for iterative optimization, while the original derives a tractable closed-form by reformulating constraints as modified rewards.

---

4. Inverse-RLignment: Inverse Reinforcement Learning from Demonstrations for LLM Alignment

URL: View paper

Brief Assessment

Inverse-RLignment[67] focuses on alignment from demonstrations using inverse RL techniques, not on deriving closed-form solutions for constrained safety optimization problems as in the original paper.

---

5. Beyond Intentions: A Critical Survey of Misalignment in LLMs.

URL: View paper

Brief Assessment

Beyond Intentions[65] is a survey paper on misalignment in LLMs. The brief mention of 'closed-form solution' in the candidate context appears to reference sorting costs, not the specific constrained-to-unconstrained reformulation with modified reward functions presented in the original paper.

---

6. Distributional preference alignment of llms via optimal transport

URL: View paper

Brief Assessment

Optimal Transport[66] focuses on distributional preference alignment using optimal transport theory for general preference data, not on deriving closed-form solutions for constrained safety alignment problems with hard safety constraints.

---

7. DarkPatterns-LLM: A Multi-Layer Benchmark for Detecting Manipulative and Harmful AI Behavior

URL: View paper

Brief Assessment

DarkPatterns-LLM[70] focuses on detecting manipulative behaviors in LLM outputs through a multi-layer benchmark framework, not on optimization formulations for safety alignment during training.

---

8. Assessing Socio-Cultural Alignment and Technical Safety of Sovereign LLMs

URL: View paper

Brief Assessment

Socio-Cultural Alignment[69] focuses on evaluating sovereign LLMs for socio-cultural alignment and technical safety across different cultural contexts, not on deriving closed-form solutions for constrained optimization problems in safety alignment.

---

9. A Survey on Training-free Alignment of Large Language Models

URL: View paper

Brief Assessment

Training-Free Alignment Survey[68] focuses on alignment methods that do not require parameter updates or fine-tuning, whereas the original paper derives a closed-form solution for a constrained optimization problem that enables direct parameter optimization. These are fundamentally different approaches to alignment.

Contribution 2: SafeDPO algorithm

Description: The authors introduce SafeDPO, a simple training method that incorporates binary safety indicators into preference optimization through a safety-aware preference transformation and an optional safety margin. It enables direct, single-stage policy updates without requiring reward models, cost models, or online sampling.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Adversarial Preference Learning for Robust LLM Alignment

URL: View paper

Brief Assessment

Adversarial Preference Learning[54] focuses on adversarial robustness through iterative adversarial training with a conditional generative attacker, while SafeDPO addresses safety alignment through preference transformation with binary safety indicators. These are distinct technical approaches to different aspects of LLM safety.

---

2. Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model

URL: View paper

Brief Assessment

Human Feedback Diffusion[57] (D3PO) focuses on fine-tuning diffusion models for image generation without reward models, while SafeDPO addresses safety-constrained preference optimization for large language models. These are fundamentally different domains (image generation vs. text generation) and problem formulations (aesthetic/alignment objectives vs. safety constraints).

---

3. A survey of direct preference optimization

URL: View paper

Brief Assessment

DPO Survey[59] is a comprehensive survey paper that categorizes and reviews existing DPO methods. While it mentions various safety-related DPO variants in its taxonomy, it does not present original algorithmic contributions that would refute SafeDPO's novelty as a specific safety-constrained preference optimization method with closed-form solutions.

---

4. Bi-factorial preference optimization: Balancing safety-helpfulness in language models

URL: View paper

Brief Assessment

Bi-Factorial Preference[52] focuses on balancing safety and helpfulness through a bi-factorial labeling function that combines two objectives (safety and helpfulness rewards) in supervised learning. SafeDPO addresses a different problem: incorporating binary safety indicators through safety-aware preference transformation to enforce hard safety constraints without requiring separate reward/cost models. The technical approaches and underlying objectives differ fundamentally.

---

5. Safety-aware preference-based learning for safety-critical control

URL: View paper

Brief Assessment

Safety-Aware Preference Learning[60] focuses on preference-based learning for safety-critical control in robotics using control barrier functions, not on single-stage preference optimization for LLM safety alignment without reward models.

---

6. Margin-aware Preference Optimization for Aligning Diffusion Models without Reference

URL: View paper

Brief Assessment

Margin-Aware Preference[53] focuses on text-to-image diffusion models and addresses reference mismatch problems, while SafeDPO targets safe language model alignment with binary safety indicators. These are fundamentally different application domains and technical approaches.

---

7. Direct preference optimization: Your language model is secretly a reward model

URL: View paper

Brief Assessment

Direct Preference Optimization[51] focuses on general preference optimization for language models without explicit safety constraints or binary safety indicators. It does not address the safety-constrained optimization problem that SafeDPO targets.

---

8. Efficient preference-based reinforcement learning via aligned experience estimation

URL: View paper

Brief Assessment

Aligned Experience Estimation[56] focuses on preference-based RL with conservative Q-function estimation and policy regularization for feedback efficiency. SafeDPO addresses safety alignment through preference transformation with binary safety indicators, which is a fundamentally different problem domain.

---

9. UniAPL: A Unified Adversarial Preference Learning Framework for Instruct-Following

URL: View paper

Brief Assessment

UniAPL[58] addresses a different problem: unified preference learning from both demonstrated and comparative preferences in a single-stage framework. SafeDPO focuses specifically on safety-constrained preference optimization with binary safety indicators, not on unifying SFT and RL modalities or resolving distributional mismatch between expert demonstrations and policy evolution.

---

10. Efficient preference-based reinforcement learning using learned dynamics models

URL: View paper

Brief Assessment

Learned Dynamics Models[55] focuses on preference-based RL using learned dynamics models for robotics tasks, not on single-stage safety-constrained preference optimization without reward models as in SafeDPO.

Contribution 3: Safety-aware preference transformation

Description: The authors propose a transformation function that reorders preference pairs based on safety indicators: safe winners remain unchanged, unsafe winners are swapped with safe losers, and unsafe-unsafe pairs are discarded. This transformation provides an unbiased estimator of the cost-augmented distribution without requiring access to the latent cost function.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

Textual similarity detection checked 21 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. A survey of direct preference optimization

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] SafeDPO: A Simple Approach to Direct Preference Optimization with Enhanced Safety View paper
- [1] Fundamental limitations of alignment in large language models View paper
- [2] Finding safety neurons in large language models View paper
- [3] Safety-aware fine-tuning of large language models View paper
- [4] Safelawbench: Towards safe alignment of large language models View paper
- [5] Safe and effective post-fine-tuning alignment in large language models View paper
- [6] Evaluating alignment in large language models: a review of methodologies View paper
- [7] Mm-safetybench: A benchmark for safety evaluation of multimodal large language models View paper
- [8] Safebench: A safety evaluation framework for multimodal large language models View paper
- [9] Ai safety in generative ai large language models: A survey View paper
- [10] Safety of multimodal large language models on images and texts View paper
- [11] Safety fine-tuning at (almost) no cost: A baseline for vision large language models View paper
- [12] Multilingual jailbreak challenges in large language models View paper

- [13] A survey of safety and trustworthiness of large language models through the lens of verification and validation View paper
- [14] Safety Misalignment Against Large Language Models View paper
- [15] Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions View paper
- [16] Navigating the safety landscape: Measuring risks in finetuning large language models View paper
- [17] Medsafetybench: Evaluating and improving the medical safety of large language models View paper
- [18] Towards comprehensive and efficient post safety alignment of large language models via safety patching View paper
- [19] Moral alignment for llm agents View paper
- [20] BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset View paper
- [21] Defining and evaluating physical safety for large language models View paper
- [22] Emerging safety attack and defense in federated instruction tuning of large language models View paper
- [23] Safe LoRA: the Silver Lining of Reducing Safety Risks when Fine-tuning Large Language Models View paper
- [24] Shaping the safety boundaries: Understanding and defending against jailbreaks in large language models View paper
- [25] Safety Alignment of Large Language Models via Contrasting Safe and Harmful Distributions View paper
- [26] Fine-tuning aligned language models compromises safety, even when users do not intend to! View paper
- [27] Unveiling the Basin-Like Loss Landscape in Large Language Models View paper
- [28] Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic View paper
- [29] Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study View paper
- [30] Fundamental Safety-Capability Trade-offs in Fine-tuning Large Language Models View paper
- [31] Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges View paper
- [32] EnchTable: Unified Safety Alignment Transfer in Fine-tuned Large Language Models View paper
- [33] Gradient surgery for safe llm fine-tuning View paper
- [34] Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack View paper
- [35] Robust-LLaVA: On the Effectiveness of Large-Scale Robust Image Encoders for Multi-modal Large Language Models View paper
- [36] Olmes: A standard for language model evaluations View paper
- [37] From Evaluation to Defense: Advancing Safety in Video Large Language Models View paper
- [38] Targeted Vaccine: Safety Alignment for Large Language Models Against Harmful Fine-Tuning via Layer-Wise Perturbation View paper
- [39] Probing the safety response boundary of large language models via unsafe decoding path generation View paper
- [40] Pku-saferlhf: Towards multi-level safety alignment for llms with human preference View paper
- [41] Towards safe large language models for medicine View paper
- [42] Equilibrate RLHF: Towards Balancing Helpfulness-Safety Trade-off in Large Language Models View paper
- [43] Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge View paper
- [44] Trustworthy llms: a survey and guideline for evaluating large language models' alignment View paper
- [45] Antidote: Post-fine-tuning Safety Alignment for Large Language Models against Harmful Fine-tuning View paper
- [46] Tuning Language Models by Proxy View paper
- [47] Trafficsafetygpt: Tuning a pre-trained large language model to a domain-specific expert in transportation safety View paper
- [48] A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More View paper
- [49] Safety Alignment Depth in Large Language Models: A Markov Chain Perspective View paper
- [50] On the safety of open-sourced large language models: Does alignment really prevent them from being misused? View paper
- [51] Direct preference optimization: Your language model is secretly a reward model View paper
- [52] Bi-factorial preference optimization: Balancing safety-helpfulness in language models View paper
- [53] Margin-aware Preference Optimization for Aligning Diffusion Models without Reference View paper
- [54] Adversarial Preference Learning for Robust LLM Alignment View paper
- [55] Efficient preference-based reinforcement learning using learned dynamics models View paper
- [56] Efficient preference-based reinforcement learning via aligned experience estimation View paper
- [57] Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model View paper
- [58] UniAPL: A Unified Adversarial Preference Learning Framework for Instruct-Following View paper
- [59] A survey of direct preference optimization View paper
- [60] Safety-aware preference-based learning for safety-critical control View paper
- [61] Decoding-time language model alignment with multiple objectives View paper
- [62] Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback View paper
- [63] Alignment of large language models with constrained learning View paper
- [64] Joint verification and refinement of language models for safety-constrained planning View paper
- [65] Beyond Intentions: A Critical Survey of Misalignment in LLMs. View paper
- [66] Distributional preference alignment of llms via optimal transport View paper
- [67] Inverse-RLignment: Inverse Reinforcement Learning from Demonstrations for LLM Alignment View paper
- [68] A Survey on Training-free Alignment of Large Language Models View paper
- [69] Assessing Socio-Cultural Alignment and Technical Safety of Sovereign LLMs View paper
- [70] DarkPatterns-LLM: A Multi-Layer Benchmark for Detecting Manipulative and Harmful AI Behavior View paper