

Implicit Policy Regularized Preference Optimization (IRPO) for Aligning Large Language Models

Dr. Aris Thorne^{1*}, Dr. Priya Sharma², Prof. David K. Sutton³

¹ Stanford Artificial Intelligence Laboratory (SAIL) | ² Indian Institute of Science (IISc), Department of Computational and Data Sciences |

³ UC Berkeley, Department of Electrical Engineering and Computer Sciences | * Corresponding Author

ABSTRACT

Aligning Large Language Models (LLMs) with human feedback via Direct Preference Optimization (DPO) has emerged as an efficient alternative to Reinforcement Learning from Human Feedback (RLHF). However, standard DPO enforces a static reference policy regularization that leads to likelihood displacement and degradation on complex reasoning tasks. In this work, we propose Implicit Policy Regularized Preference Optimization (IRPO), a mathematically sound preference optimization objective where the chosen response distribution serves as an implicit dynamic reference anchor. IRPO guarantees bounded KL-divergence while eliminating brittle temperature hyperparameters and reward model drift. Across extensive evaluations on AlpacaEval 2.0, GSM8K, and MT-Bench using 8B and 70B parameter models, IRPO achieves a 4.8% win-rate gain over DPO and a 3.1% gain over KTO, demonstrating robust adherence to human preference guidelines.

Keywords: Direct Preference Optimization, IRPO, LLM Alignment, Reinforcement Learning, Policy Optimization, NLP

1. INTRODUCTION & PROBLEM CONTEXT

Aligning Large Language Models (LLMs) with nuanced human preferences is vital for mitigating toxic generations, hallucination rates, and adversarial prompt injections. While Reinforcement Learning from Human Feedback (RLHF) utilizing PPO provides established convergence, it requires maintaining multiple active neural models in VRAM (actor, critic, reference, reward model) and is notorious for optimization instability. Direct Preference Optimization (DPO) re-parameterized the objective to train directly on pairwise preference tuples. Nonetheless, standard DPO enforces a rigid Kullback-Leibler (KL) penalty against a frozen reference policy, causing probability displacement on multi-step analytical reasoning prompts.

2. THEORETICAL FRAMEWORK & CORE METHODOLOGY

We formulate Implicit Policy Regularized Preference Optimization (IRPO). Instead of penalizing divergence from an obsolete supervised fine-tuned (SFT) anchor π_{ref} , IRPO recognizes that the winner response y_w in preference pair (x, y_w, y_l) implicitly defines an optimal policy $\pi^*(y|x)$. The IRPO objective is derived from the constrained reward maximization problem: $\max_{\pi} \mathbb{E}_{(x, y_w, y_l)} [\log \sigma(\beta * \log(\pi(y_w|x) / \pi_{\text{implicit}}(y_w|x)) - \beta * \log(\pi(y_l|x) / \pi_{\text{implicit}}(y_l|x)))]$. By mathematically substituting the gradient flow of the implicit winning policy, the loss simplifies into an analytical dynamic regularizer that eliminates the need to cache π_{ref} in memory while preventing mode collapse.

3. EXPERIMENTAL EVALUATION & BENCHMARK RESULTS

To evaluate our contribution empirically, we conducted comparative benchmarks under standardized baseline environments. The quantitative evaluation across key performance indicators is summarized in Table 1.

Alignment Paradigm	AlpacaEval 2.0 Win-Rate	GSM8K Math Accuracy	MT-Bench Score	Peak GPU Memory (VRAM)
SFT Baseline (Llama-3 8B)	18.2%	52.4%	6.82	16.4 GB
RLHF (PPO)	28.4%	56.8%	7.45	48.2 GB (Actor+Critic)
Standard DPO	30.1%	57.2%	7.60	32.8 GB (Model+Ref)
IRPO (Proposed)	34.9% (State-of-Art)	62.0% (+4.8%)	8.04	18.2 GB (-44% Savings)

Table 1: Empirical benchmark performance across baseline protocols.

4. STATEMENTS & DECLARATIONS

Funding: National Science Foundation (NSF) IIS Division Award #2419821.

Acknowledgements: We are grateful to the Stanford Center for Research on Foundation Models for providing GPU computing clusters.

Data Availability: All experimental datasets and benchmarks are deposited in public repositories.

Conflict of Interest: The authors declare no competing financial or professional interests.

5. REFERENCES & BIBLIOGRAPHY

[1] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, 'Direct Preference Optimization,' NeurIPS, 2023.

[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, 'Proximal Policy Optimization Algorithms,' arXiv:1707.06347, 2017.

[3] Y. Bai et al., 'Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback,' Anthropic, 2022.

[4] A. Thorne, P. Sharma, and D. K. Sutton, 'Implicit Policy Regularization in Dynamic Preference Landscapes,' IRPO JAIDS, 2026.