

Scaling Down AI Safety: Evaluating RLAIF for Small Language Models

Will Fowler
Akash Chatterji

1 Introduction

As Artificial Intelligence (AI) becomes more powerful and widely adopted, ensuring that they behave safely and align with human values is becoming one of the more pressing challenges that is present in this field today. Large Language Models (LLMs), systems which generate text, answer all types of questions and assist with everyday tasks can sometimes produce harmful or misleading content. To reduce such risks, leading AI labs use *Reinforcement Learning from Human Feedback (RLHF)* to guide models towards safer behavior. However, this process is costly and resource-intensive, requiring large teams of human reviewers and powerful computing infrastructure and putting it out of reach for most academic and independent researchers. This led to the development of a more accessible alternative: *Reinforcement Learning from AI Feedback (RLAIF)*, where AI systems themselves provide the feedback, guided by clear constitutional principles that define desirable behavior.

By applying this method to models on the smaller side, we aim to test whether meaningful safety improvements can be achieved with limited resources. We will compare two methods for training models to adhere to human preferences: Proximal Policy Optimization (PPO) and Direct Policy Optimization (DPO). This approach could help democratize AI safety research, allowing more people to experiment, learn, and contribute to building safer, more aligned AI systems.

2 Background Related Work

2.1 Motivation

Large language models (LLMs) are increasingly deployed in real-world applications, making their alignment with human values a critical concern. Reinforcement Learning from Human Feedback (RLHF) has emerged as the dominant paradigm for aligning LLMs, but it faces a significant scalability challenge: the need for extensive human labeling is both expensive and time-consuming. Since this is a class project, we will be working with much smaller LLMs than those on

the frontier. This means we can use more powerful models as a source of preference data instead of humans, so that smaller models can be aligned with a much lower cost and at a greater scalability. Within the RLHF/RLAIF framework, two primary optimization methods have gained prominence: Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO). While PPO has been widely adopted in practice, DPO has recently attracted attention for its theoretical simplicity and computational efficiency, as it eliminates the need for a separate reward model. Xu et al. [5] directly compare these methods in the RLHF setting, finding that DPO has some limitations. However, the relative performance of PPO and DPO in the RLAIF setting remains underexplored. Our project aims to explore the relative performance of PPO and DPO in the RLAIF setting by using AI-feedback to align the LLaMa3.1 8b LLM.

2.2 Related Work

- **RLAIF Foundations:** Bai et al. (2022) introduced Constitutional AI, demonstrating that AI feedback can effectively replace human preferences for alignment. Their framework uses a "constitution"—a set of principles encoded in natural language—to guide an AI critic in evaluating model outputs.
- **Challenges in RLHF:** Casper et al. (2023) provide a comprehensive review of RLHF challenges, highlighting the cost-performance trade-off inherent in human annotation.
- **PPO for LLM Alignment:** Proximal Policy Optimization (Schulman et al., 2017) has become the de facto standard for RLHF implementation. The typical pipeline involves: (1) supervised fine-tuning of a base model, (2) training a reward model on preference data, and (3) using PPO to optimize the policy model against the reward model while maintaining proximity to the original model.
- **DPO as an Alternative:** Direct Preference Optimization (Rafailov et al., 2023) reformulates the RLHF objective as a supervised learning problem. Rather than learning an explicit reward function, DPO directly optimizes the policy to satisfy through a classification-style loss on preference pairs.
- **PPO vs DPO:** Xu et al. (2024) compare the performances of using DPO and PPO for RLHF in LLMs. They find that, while DPO outperforms PPO in many academic benchmark settings, it has limitations on out-of-distribution inputs.

3 Technical Approach / Methodology / Theoretical Framework

For this project, we aim to fine-tune language models using reinforcement learning (RL) techniques that mainly incorporate preference-based feedback. Our primary method will be **Proximal Policy Optimization (PPO)** [6], an RL algorithm that updates the model through reward-driven policy optimization. As referenced above, to evaluate the effectiveness and efficiency of this RL approach, we will also include **Direct Preference Optimization (DPO)** [4] as a non-RL comparative baseline. Both methods optimize a policy $\pi_\theta(y|x)$ which represents the model's conditional distribution over responses y given input x towards behaviors that better align with human or AI-generated preference signals.

3.1 Proximal Policy Optimization (PPO)

PPO is a policy gradient method that improves stability by limiting the size of each policy update. Its objective is defined as:

$$L_{PPO}(\theta) = E_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (1)$$

where:

- $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}$ represents the ratio of how likely the new policy chooses action a_t compared to the old policy. So essentially it measures how much the policy has changed.
- \hat{A}_t is the advantage estimate, which tells the model how much better or worse an action performed compared to what was expected.
- ϵ is a small clipping value that prevents the policy from changing too much in one update, helping keep training overall stable.

So in simpler terms, PPO will only update the policy if the change improves performance *without moving too far* from the previous version, ensuring stability and preventing over-optimization.

3.2 Direct Preference Optimization (DPO)

In contrast to PPO, DPO reformulates policy optimization as a direct likelihood-based problem, removing the need for explicit rewards and reinforcement learning loops. Its objective is defined as:

$$L_{DPO}(\theta) = -E_{(x,y^+,y^-)} \left[\log \sigma \left(\beta \left(\log \frac{\pi_\theta(y^+|x)}{\pi_{ref}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{ref}(y^-|x)} \right) \right) \right] \quad (2)$$

where:

- (x, y^+, y^-) are the examples of inputs x with two possible outputs: y^+ (the preferred response) and y^- (the less preferred response).
- π_{ref} is the reference policy, usually the original supervised fine-tuned model, which acts as a baseline for comparison.
- σ is the sigmoid function, which converts a difference in scores into a probability between 0 and 1.
- β is used as the scaling factor that determines how strongly the model should prioritize the preferred response over the less preferred one.

Overall, DPO aligns the model directly with preferred outputs by increasing the likelihood of y^+ over y^- , offering a simpler and more efficient alternative to PPO and using it as a strong comparative baseline.

4 Experiments and Evaluation

After training with both PPO and DPO, we will have two fine-tuned models to compare against the base LLaMA 3.1 8B-Instruct model.

4.1 Qualitative Assessment

This will be manual observation of model outputs to assess alignment and general usability. Our observations will be described, qualitatively, in our final report.

- Models tested: Base LLaMA 3.1 8B-Instruct, DPO fine-tuned model, PPO fine-tuned model
- Test scenarios:
 - Simple question answering (factual queries across various domains)
 - Thought experiments and hypothetical reasoning
 - Adversarial prompts testing for harmful outputs or misalignment
- We will sample around 20 different prompts for each model and grade the responses on alignment and whether their answer is correct/expected

4.2 Quantitative Assessment

We aim to demonstrate two key outcomes: (1) improved alignment to human values, and (2) preservation of general capabilities. These results will be conveyed as plots in our final report.

- Alignment & Safety Benchmarks:
 - TruthfulQA: Measures truthfulness and resistance to generating common misconceptions

- PKU-SafeRLHF: Tests for harmful content generation across multiple risk categories
- Capability Benchmarks:
 - MMLU: Measures general knowledge and reasoning across 57 subject areas
 - HellaSwag: Tests commonsense reasoning and prediction abilities

4.3 Backup Plan

Measuring "alignment to human values" can be challenging, and since we are working with relatively small models, there is a chance that we will not observe an effect on alignment benchmarks. This might occur because the stock LLaMa 3.1 8B model is already well-aligned from instruction tuning so improvements will be hard to detect. In this scenario, we can still answer our main research question (how does DPO perform compared to PPO in RLAIF?) by modifying what we aim to align the model toward. In this case, we will rerun the experiment with a more objectively measurable target behavior. Specifically, we will use RLAIF to train the model to respond in a strict structured format, such as "Answer: [response] — Confidence: [high/medium/low] — Reasoning: [brief explanation]" Then, we will be able to observe the effects much more easily since evaluation can be as simple as checking if the model responds in the desired format or not. This would still answer our project's core question and serve as a proof-of-concept for using RLAIF on larger models where the alignment effect can be observed.

5 Timeline and Individual Responsibilities

5.1 Timeline

- Week 1:
 - Generate AI preference dataset using API model for feedback on LLaMA 3.1 8B outputs
 - PPO planning phase: Try to find resources. Figure out what libraries are required.
- Week 2:
 - Go through Unslloth DPO notebook with AI-generated preference data
 - Begin work on PPO training pipeline
- Weeks 3-4:
 - Continue and Finish PPO pipeline

- Start evals on DPO and base model
- Week 5:
 - Run benchmarks on all three models: base, DPO, PPO
 - Conduct qualitative evaluation
- Week 6:
 - Work on report
 - Extra time for backup plan

5.2 Individual Responsibilities

- Person 1:
 - Research tutorials/notebooks for PPO
 - Work on PPO training loop
- Person 2:
 - Generate AI-perference dataset on base model outputs
 - Use Unslloth notebook to train the DPO model
- Shared Responsibilities:
 - Running evaluations on base LLaMa model whenever they have extra time
 - Running quantitative benchmarks
 - Qualitative evaluations
 - Writing the report

References

- [1] Sutton, R. S., and Barto, A. G. *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press, 2018.
- [2] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [3] Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2307.15217*, 2023.

- [4] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- [5] Xu, S., Fu, W., Gao, J., Ye, W., Liu, W., Mei, Z., Wang, G., Yu, C., and Wu, Y. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. *arXiv preprint arXiv:2404.10719*, 2024.
- [6] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [7] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems*, 30, 2017.