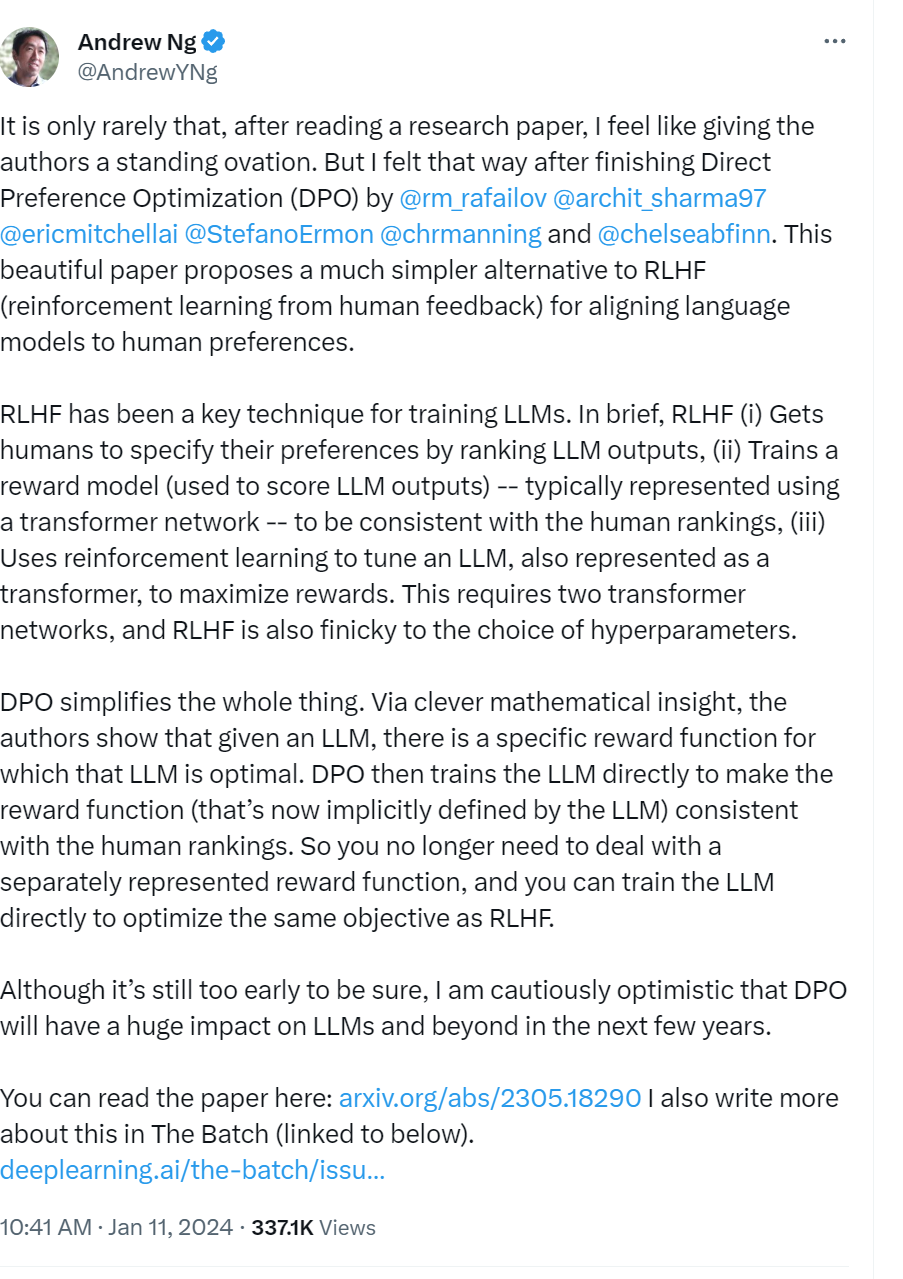
## Reading Group 1/14/2023 Direct Preference Optimization

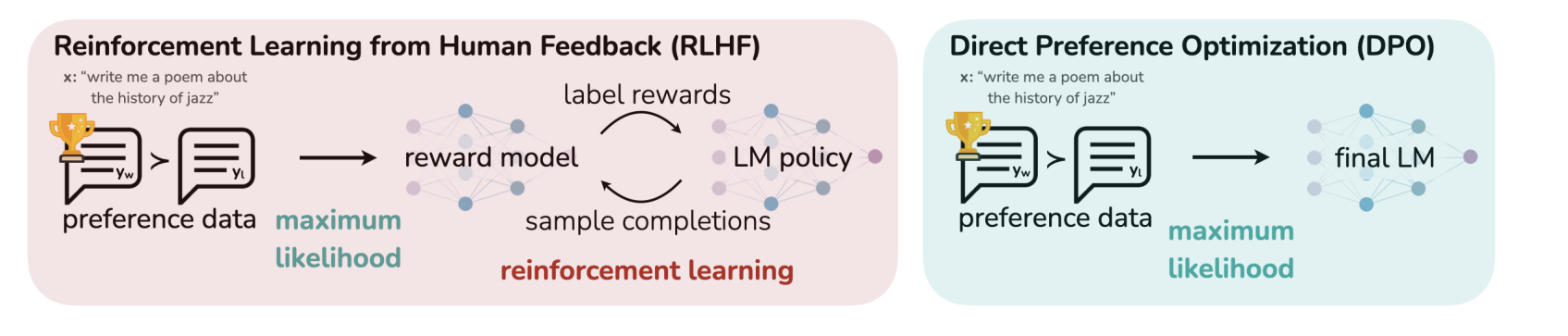
## Paper

* <https://arxiv.org/abs/2305.18290>

## Inspiration behind this reading



## Introduction

* “Similarly, we might want our language model to be aware of a common misconception believed by 50% of people, but we certainly do not want the model to claim this misconception to be true in 50% of queries about it!”
  + Tldr says that current rlhf methods instill unwanted behaviors in LLMs
* Claims to improve upon RLHF using a simple binary cross entropy objective
* They show how to directly optimize a language model to adhere to human preferences, without explicit reward modeling or reinforcement learning.
* 
* Similar to RLHF in that it is reward maximization with KL divergence constraint
  + Claims DPO is simpler to implement
* Mathematical stuff
  + DPO update increases relative log probability of preferred to dispreferred responses
  + Dynamic per example importance weight
    - This prevents model degeneration
* Author explained it as the closed form
* Claims it is just as effective as existing methods, like RLHF

## RLAIF

* <https://arxiv.org/abs/2309.00267>

## 

## 

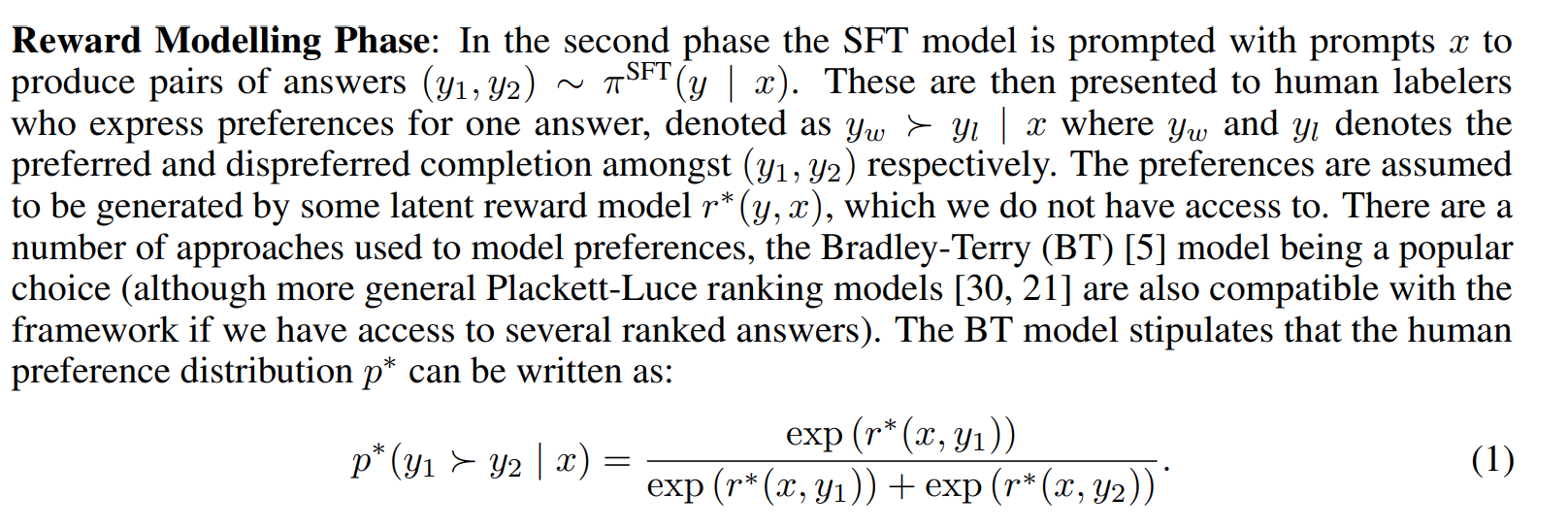
## Related Work

* Instruction-tuning can improve performance
  + Methods like these first optimize a neural network reward with a preference dataset
  + then fine-tune a language model to maximize the given reward using reinforcement learning algorithms
* Learning from preferences has also been studied in contextual bandits
  + Contextual bandit learning using preferences or rankings of actions
    - Aka contextual dueling bandit
* Preference based RL
  + Learns from binary preferences generated by an unknown scoring function

## 

## RLHF pipeline

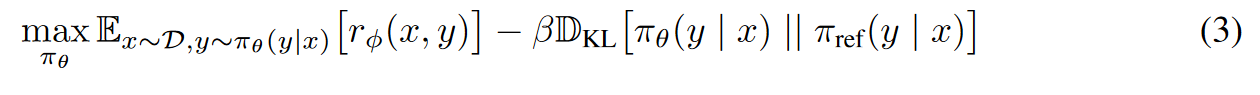
* Has 3 phases
  + 1. Supervised Fine tuning
  + 2. Preference sampling and reward learning
  + 3. RL optimization



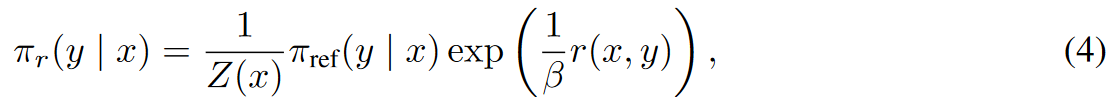
## 

## Direct preference optimization

* Doesnt learn a reward and optimize via RL
* Instead, uses reward model parameterization that lets us use the closed form of the optimal policy
  + Doesn’t use rl in a loop



Gets derived into



## Tldr

* Learning from preferences is scalable and powerful for training intelligent language models
* introducing DPO
  + For training language models from preferences **without reinforcement learning**
  + Rather than coercing the preference learning problem into a standard RL setting in order to use off-the-shelf RL algorithms, DPO identifies a mapping between language model policies and reward functions that enables training a language model to satisfy human preferences directly
  + with a simple cross-entropy loss, without reinforcement learning or loss of generality.
  + With virtually no tuning of hyperparameters, DPO performs similarly or better than existing RLHF algorithms, including those based on PPO; DPO thus meaningfully reduces the barrier to training more language models from human preferences

## <https://towardsdatascience.com/understanding-kl-divergence-f3ddc8dff254>

## 

## Limitations and future work

* Does DPO generalize out of distribution compared with learning form an explicit reward function ?
  + Suggests that DPO **can**  generalize similarly to PPO based models, but more study is needed
  + Can training with self labeling form the dpo policy make effective use of unlabeled prompts ?
* How does reward over optimization manifest in the DPO setting and is slight decrease in performance in figure 3 an instance of it ?
* Also they want to scale up DPO to SOTA models
* Does dpo work with training models in other modalities?

## 

## Code

* <https://github.com/eric-mitchell/direct-preference-optimization>
* Datasets: <https://github.com/eric-mitchell/direct-preference-optimization/blob/main/preference_datasets.py>
* 