https://arxiv.org/pdf/2305.18290.pdf

Background

Selecting model's desired responses and behavior form wide knowledge and abilities is crucial
Precise control of behavior of LMs is hard due to unsupervised nature of training

- Existing methods collect human labels of relative quality of model generations and fine tune LM to align with preference (RLHF)
- RLHF is complex and unstable
  - First fit reward model to reflect human preferences (a dataset of prompts and human preferences over pairs of responses) and then fine tune LM using RL to maximize estimated reward without drifting too far from original model (find a policy)
  - Objective: reward maximization with KL divergence constraint

Despite success of instruction tuning, relative human judgments of response quality are often easier to collect than expert demonstrations -> subsequent works have fine-tuned LLMs with datasets of human preferences

Idea

New parameterization of reward model in RLHF to extract optimal policy in closed form (fitting implicit reward model)

- Solve RLHF problem with simple classification loss
- Stable, performant, computationally lightweight
- Eliminates need for sampling from LM during fine-tuning or during significant hyperparameter tuning
- Increases relative log probability of preferred to dispreferred response but incorporates a dynamic, per- example importance weight that prevents model degeneration that we find occurs with naive probability ratio objective
  - Relies on theoretical preference model (Bradley Terry model) that measures how well a given reward function aligns with empirical preference data
  - Change of variable to define preference loss as a function of the policy directly

Mapping from reward functions to optimal policies, transform a loss function over reward functions into a loss function over policies

Control sentiment of generations
Improves response quality in summarization and single turn dialogue