Title: Reinforcement Learning with Human Feedback (RLHF): Bridging the Gap between AI and Human Expertise

Proposal: This project explores the integration of human feedback into reinforcement learning algorithms, aiming to enhance AI systems' decision-making capabilities. Key papers examined include "*Secrets of RLHF in Large Language Models Part I and II*" by Rui Zheng et al. This project consists of code implementation of fine-tuning an LLM using RLHF and evaluating it based on various metrics aligning with the results presented in these papers to understand the impact of human feedback on reinforcement learning models, paving the way for more effective and collaborative AI systems.

Paper

Secrets of RLHF in Large Language Models, Secrets of RLHF in Large Language Models Part I: PPO https://arxiv.org/pdf/2307.04964.pdf

Secrets of RLHF in Large Language Models, Part II: Reward Modeling https://arxiv.org/pdf/2401.06080.pdf

Code Repo and Tutorials

https://learn.deeplearning.ai/reinforcement-learning-from-human-feedback

https://medium.com/@madhur.prashant7/rlhf-reward-model-ppo-on-llms-dfc92ec3885f

https://github.com/HumanSignal/RLHF/blob/master/tutorials/RLHF_with_Custom_Datasets.ipynb

https://github.com/opendilab/awesome-RLHF?tab=readme-ov-file#2024