Reinforcement learning and Human Feedback (RLHF)

1. Lower the toxicity of the LLM finetuned.
2. PEFT, TRL (PPO trainer)
3. AutoModelForSeq1 classification- used to load the Facebook binary classifier (RoBERTaseq classifier), which say weather or not a prompt is a hate speech
4. Used PEFT as a baseline model for RLHF
5. PPO model parameter to be update= Value\_Head+769 parameters (768dimension of Value Head+1bias).
6. Train – reference model and current PPO model. Then compared using KL divergence.( and minimizes the models ability to perform the reward hacking)
7. In PPO training, the optimizations is done through logits and not through SoftMax probabilities.(specifically Value Heads will be updated)
8. Sentiment pipeline.
9. Goal: to reduce mean toxicity score after we do PPO.
10. PPO trainer used.
11. Reference model in previous folder (Lora\_PEFT), used for KL divergence during PPO training. This KL divergence is what is intended to minimize.
12. KL-divergence high- reward hawk
13. Gradient updating through PEFT.
14. There KL divergence converges around 27