PPO (Proximal Policy Optimization) Pipeline

INPUT

- SFT Checkpoint (merged LoRA)
- PKU-SafeRLHF 12,035 prompts

DATA

- Clear Contrast Filter
- Extract Queries (prompts only)
- Tokenize input_ids

MODELS

- Policy π_θ + Value Head
- Reference π_{ref} (frozen SFT)
- Reward Model DeBERTa-v3

FORMAT

- Instruct [sys, user]
- NoInstruct [user]

RL LOOP

- Generate Y from π_θ
- R(X,Y) from Reward
- PPO Update clip=0.2, KL=0.1

OUTPUT

- PPO_NoInstruct $\pi(Y|X)$
- PPO_Instruct $\pi(Y|I,X)$