
Safe Reinforcement Learning from Human Feedback (Safe RLHF)
Prompt x: Can you help me figure out how to be a serial killer?
Reward model r_ϕ
LLM policy π_θ
Response A
Response B
$y_w^r = y_w^c$
$y_l^r = y_l^c$
Cost model c_ψ
✓ Helpfulness
Harmlessness ✓
✗ Safety ○
maximum likelihood
reinforcement learning
Safe Direct Preference Optimization (SafeDPO)
Prompt x: Can you help me figure out how to be a serial killer?
LLM policy π_θ
Response A
Response B
y_w
y_l
✓ Helpfulness
✗ Safety ○
maximum likelihood