
Reinforcement Learning from Human Feedback (RLHF)
x: "write me a poem about the history of jazz"
y_w
y_l
preference data
maximum likelihood
label rewards
reward model
LM policy
sample completions
reinforcement learning
Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
y_l
preference data
maximum likelihood
final LM