Reinforcement learning cont

* Evaluation (after training measure performance)

1) Freeze the policy (no more learning)

2) run agent many episodes in the enviroment

3) collect metrics like:

- Avg reward per episode
- success rate (eg % of times robot reaches goal)
- Cumulative reward curve over time
- stability (does performance vary alot)

- this tells us how well the agent generalize not just memorized

RL vs RLHF

RL:
- Env: game/robtic world
- Agent: learns by taking action in env
- Reward: comes directly from env, Ex +1: win -1 Lost
- training loop: Agent in env → reward → update policy
- Eval: Avg score, success rate

-used in games, cars, robtics

RLHF:
- Env: human preference data
- Agent: a LLM
- Reward: not built in → learned from humans

1) collect LLM outputs
2) humans Rank which one's better
3) train Reward model to predict human preference
4) use RL (PPO Actor method) to fine tune LLM for maximized reward

- Training: LLM generate → reward model scores → policy update
- Eval: human eval, quality checks, alignment benchmarks

-used in ChatGPT, Claude etc