The Reinforcement Learning Trap

How current RL approaches extract minimal signal from massive computation

Generate hundreds of solution attempts in parallel → Check which attempts produced correct final answers → Upweight every step in successful trajectories

1000s

Steps in long reasoning rollout

1 bit

Information extracted (right/wrong)

100%

Steps reinforced in success