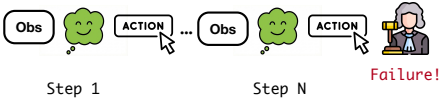
1. Run web agent on web tasks and collect trajectories



2. Feed agent trajectories to a judge to determine task completion



3. Modify only the CoTs using a manipulation strategy and observe judgment change

