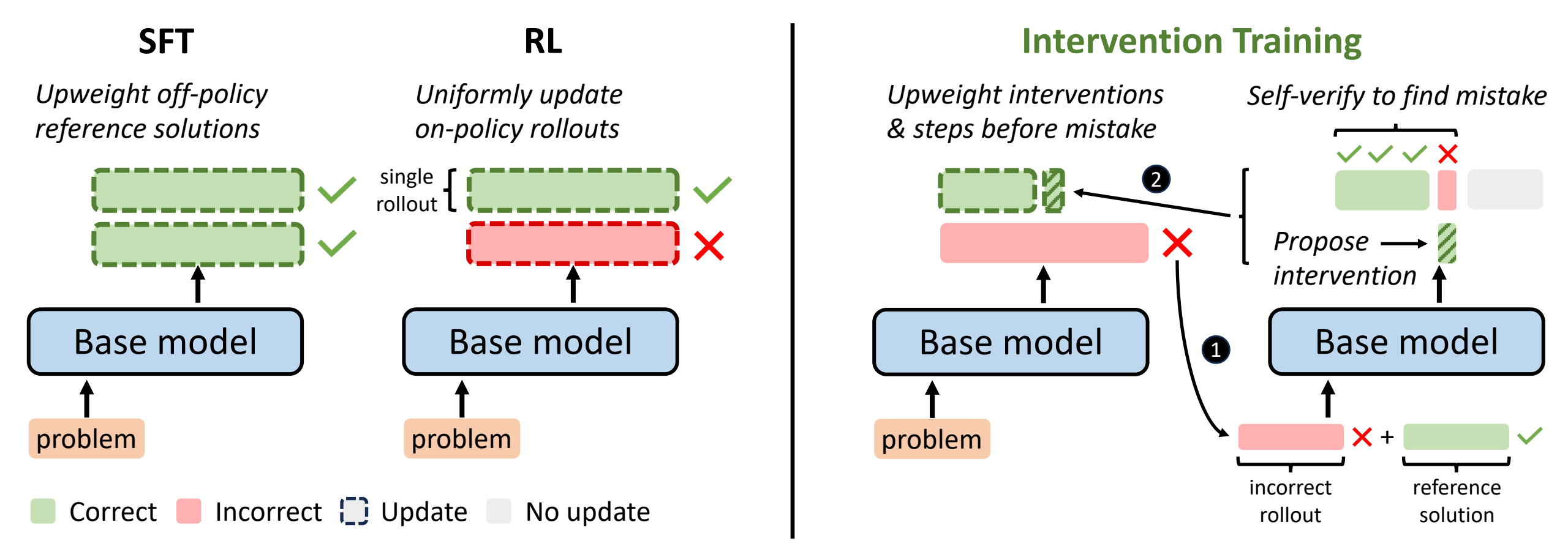
SFT
RL
Upweight off-policy reference solutions
Uniformly update on-policy rollouts
single rollout
Base model
Base model
problem
problem
Correct
Incorrect
Update
No update
Intervention Training
Upweight interventions & steps before mistake
Self-verify to find mistake
2
Propose intervention
Base model
Base model
1
problem
incorrect rollout
reference solution