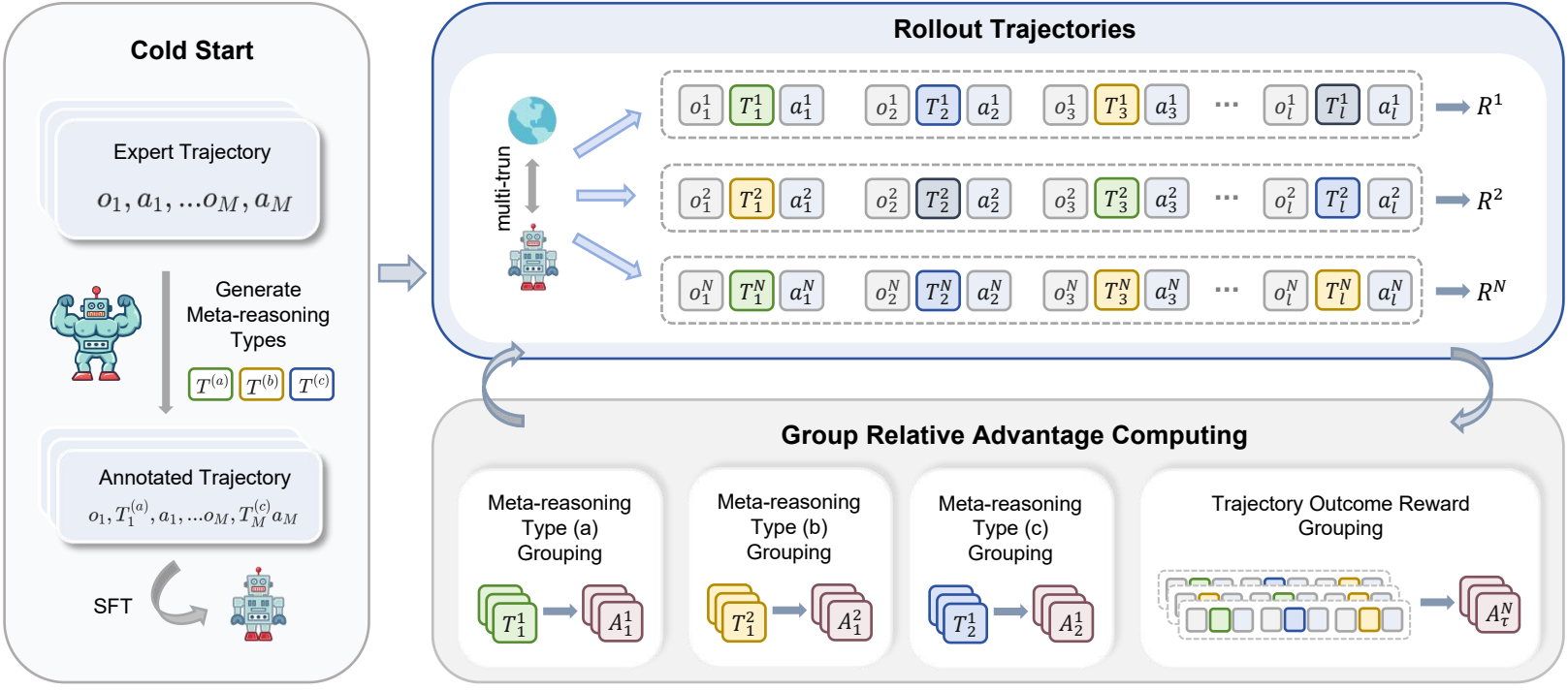
Cold Start
Expert Trajectory
$o_1, a_1, ...o_M, a_M$
Generate Meta-reasoning Types
$T^{(a)}$ $T^{(b)}$ $T^{(c)}$
Annotated Trajectory
$o_1, T_1^{(a)}, a_1, ...o_M, T_M^{(c)} a_M$
SFT
Rollout Trajectories
multi-trun
R^1
R^2
R^N
Group Relative Advantage Computing
Meta-reasoning Type (a) Grouping
Meta-reasoning Type (b) Grouping
Meta-reasoning Type (c) Grouping
Trajectory Outcome Reward Grouping