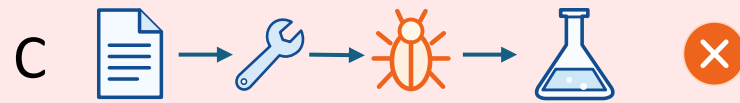
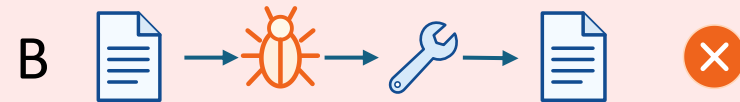
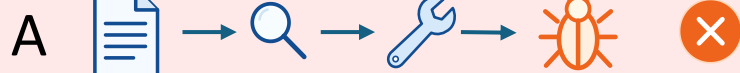


1. Rollouts



Task: Fix the bug in the repository



2. Binary Verifier



Run tests/verifier

0

0

0

Same outcome,
Different behavior

1

Pass Test

3. Behavioral Self-Rewarding

Inputs



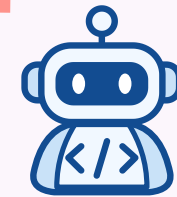
Task



Trajectories



Ground-truth



Same Policy
Model

Behavior Rubrics

1. Targeted actions

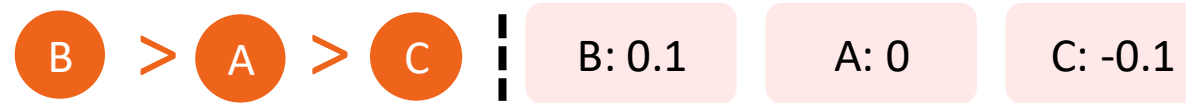
2. Use of valid tools

3. ...

Pairwise comparison



Ranking and refined rewards



4. Final Rewards

$$R = R_{ver}(\tau_i) + R_{BSR}(\tau_i)$$

	R_{ver}	R_{BSR}	R
A	0	0.0	0.0
B	0	0.1	0.1
C	0	-0.1	-0.1
D	1	0.0	1.0