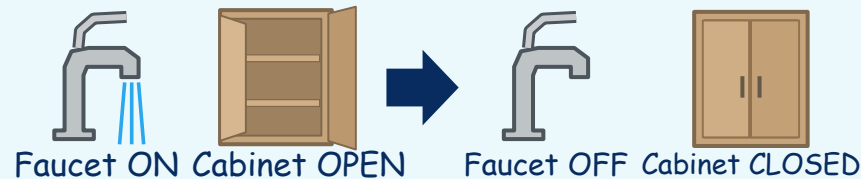


1. Task Template

Task specification

Required objects: faucet, cabinet
Initial: faucet ON; cabinet OPEN
Goal: faucet OFF; cabinet CLOSED

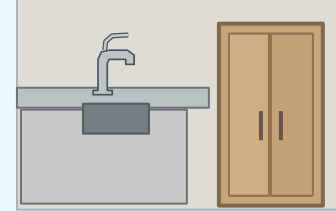


Optional Failure / Perturbation

⚡ Trigger: first Close of Cabinet
Effect: cabinet remains partly open

2. Scene Setup

1 Match object categories



- ✓ Faucet found
- ✓ Cabinet found

2 Initialize scene



- Set required states
- Add objects as needed

Bound task instance

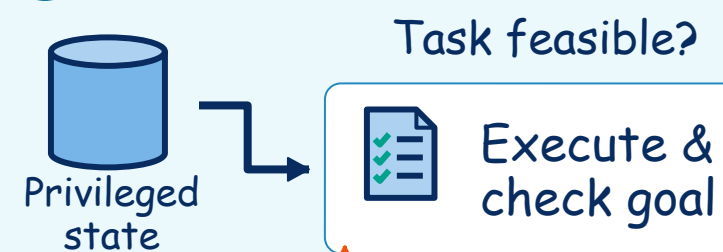
Inject at matched operation

3. Task Validation

1 Symbolic Checks

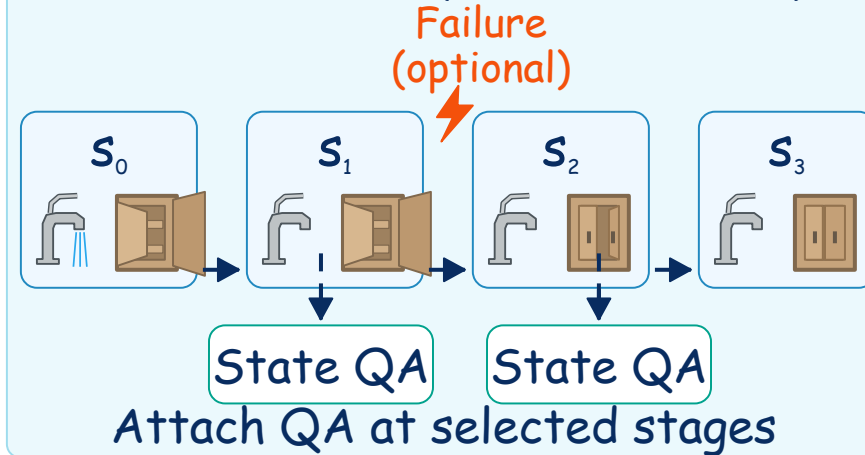


2 Oracle Execution Check



4. Trace & QA

Execution trace (validated task)



Validated Tasks → Traces + State-linked QA

Shared Benchmark Infrastructure

Rule-Based Oracle Execution



Exploration



Navigation



Manipulation

calls

Execution calls

Model Evaluation



Model

Answers



State-linked QA

Questions