Task ID: _____

Description

Describe the task, the capability it measures, its environment, and the initial assets provided to the agent.

Construction

Explain how you constructed the task and which verification techniques you used.

Outcomes

Metric	Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5
Verifier reward					
Wall-clock time					
Tokens or cost					

Summarize the most informative success, failure, or difference between models.

Self-assessment

Axis	Rating (1–5)	Brief justification
Reproducibility		
Capability		
Task–verifier fidelity		
Difficulty		