Review 1: `<task-id>`

Summary

In one or two sentences, what capability does this task measure?

Ratings

Criterion	Rating (1–5)	Justification
Reproducibility	–	*One sentence.*
Capability	–	*One sentence.*
Task–verifier fidelity	–	*One sentence.*
Difficulty	–	*One sentence.*
Report	–	*One sentence.*

Evidence

Support your ratings by citing at least one specific task file, verifier check, or trajectory event.

Strongest aspect

One or two sentences.

Suggested improvement

One concrete, actionable change, in one or two sentences.

Review 2: `<task-id>`

Summary

In one or two sentences, what capability does this task measure?

Ratings

Criterion	Rating (1–5)	Justification
Reproducibility	–	*One sentence.*
Capability	–	*One sentence.*
Task–verifier fidelity	–	*One sentence.*
Difficulty	–	*One sentence.*
Report	–	*One sentence.*

Evidence

Support your ratings by citing at least one specific task file, verifier check, or trajectory event.

Strongest aspect

One or two sentences.

Suggested improvement

One concrete, actionable change, in one or two sentences.

Review 3: `<task-id>`

Summary

In one or two sentences, what capability does this task measure?

Ratings

Criterion	Rating (1–5)	Justification
Reproducibility	–	*One sentence.*
Capability	–	*One sentence.*
Task–verifier fidelity	–	*One sentence.*
Difficulty	–	*One sentence.*
Report	–	*One sentence.*

Evidence

Support your ratings by citing at least one specific task file, verifier check, or trajectory event.

Strongest aspect

One or two sentences.

Suggested improvement

One concrete, actionable change, in one or two sentences.