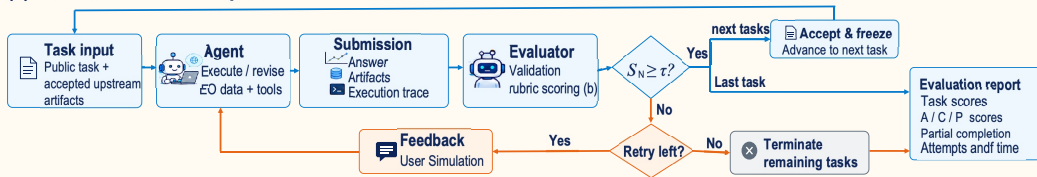


(a) Multi-turn evaluation protocol



(b) Rubric-based task evaluation

Evidence inputs

Private references

Reference answers
Reference artifacts

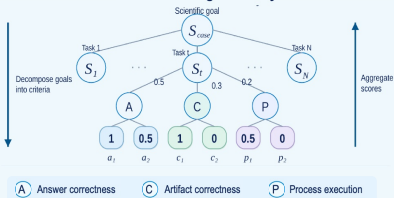
Submission evidence

answer.md
Artifacts
Execution trace

To judging
and scoring

Private references are hidden
from the agent.

Frozen scoring hierarchy



Selected scientific criteria

Longbasaba glacier-lake example

Criterion	Evidence check	Completion
(a1) Lake-area change	numeric check	1
(c1) Coverage, 1988–2018	CSV content check	1
(p1) Trace–output linkage	Trace check	0.5

LLM judgment + programmatic checks

P measures trace-based evidence coverage.