

Scientists work entirely inside OpenReview; the harness runs everything else and replies in the same thread.

A Multi-round peer review

Proposer and nominated reviewer iterate until "It is acceptable."

B Task upload

Harbor-format ZIP posted with /submit_task

C Preflight validation

Manager approves; the harness checks the package and environment

D Agent execution

Proposer sends /run_agent; three frontier agents each write a paper and artifacts

E Scientist feedback

Proposer critiques each paper, ranks them and revises the rubric; /run_agent starts the next round

OpenReview.net

PDF

by Task Proposer SciVeri-Bench 2026 Internal Review

Task Identifier: Domain: Evaluation Metric Adequacy:

Official Review · after multi-round discussion

Official Review by Task Reviewer

Rating 8 Novelty 4/5 Significance 5/5 Open-ended yes "It is acceptable."

Task upload

Task Upload by Task Proposer

Task Zip: zip Harbor format Comment: /submit_task

Official Comment by Task Manager

/validate_task

Task preflight validation passed

AI Outcome by SciVeri Harness

Package Environment

Official Comment by Task Proposer

/run_agent

Round 1 · Papers and scientific feedback

AI Outcome by SciVeri Harness

A Paper A — Claude Code Opus 5.5 PDF ZIP

B Paper B — Codex GPT-6-Astra PDF ZIP

C Paper C — OpenScience GPT-6-Astra PDF ZIP

Scientific Weakness Scientific weaknesses — Paper A by Task Proposer

Pairwise Ranking Scientific taste — pairwise ranking by Task Proposer

Rubric Revision Updated task rubric by Task Proposer

Round 2 · Papers and scientific feedback

AI Outcome by SciVeri Harness

WHO POSTS IN THE THREAD

- Task Proposer

domain scientist
Uploads the task, launches /run_agent, gives feedback each round
- Task Reviewer

nominated expert
Reviews over multiple rounds until "It is acceptable."
- Task Manager

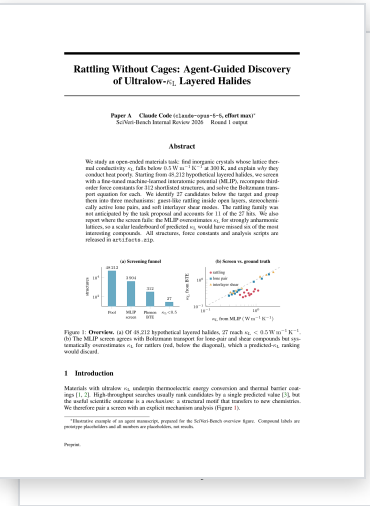
AI researcher
Approves the task with /validate_task; helps with Harbor code
- SciVeri Harness

automated
Validates the task, runs the agents, posts papers & feedback forms

Round 1 · AI Outcome

Paper A — Claude Code

claude-code, claude-opus-5-5 effort max



PDF paper.pdf

Rattling Without Cages: Agent-Guided Discovery of Ultralow-κ_L Layered Halides

NeurIPS 2026 style · [preprint] · 3 pages

ZIP artifacts.zip

structures/ · force_constants/ · analyze.py · results.csv

Scientific Weakness

Scientist feedback · every round

Reply forms → verification tasks

- Scientific weaknesses

per paper, with evidence → Critique Generation
- Pairwise ranking

A B Tie

per pair → Taste Evaluation
- Updated task rubric

complete new rubric.md → Rubric Generation

Reply /run_agent → Round 2 with the new rubric