KDD 2026 WORKSHOP · EVALUATION AND TRUSTWORTHINESS OF AGENTIC AI

Accepted paper #15
Jeju, Korea · August 9, 2026

MacCode Lab: A Reproducible Single-Machine Harness for Execution-Guided Code Agents, and When Repair Actually Helps

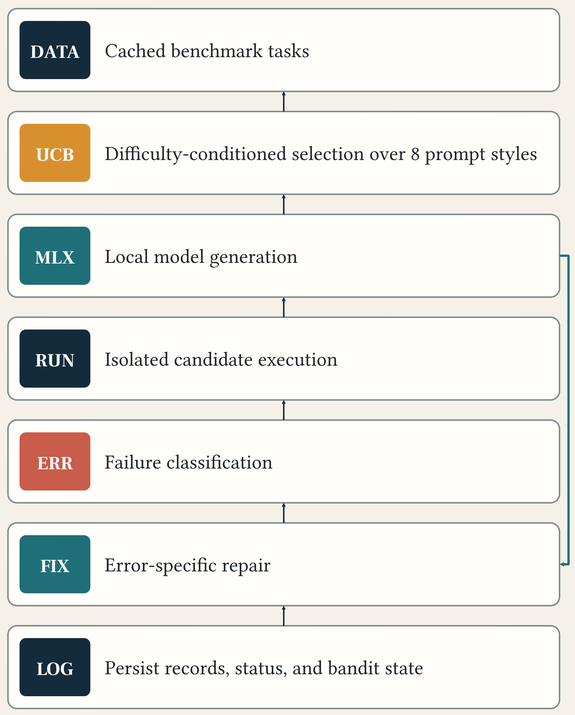
Mayank Sharma · Rohit Kumar Mourya
Indian Institute of Technology Jodhpur, India
b23es1023@iitj.ac.in · b23es1029@iitj.ac.in

Repair depends on the model. | **Public-only scoring can change the comparison.** | **One workstation can run and replay the full loop.**

Question

Can a single workstation run an auditable code-agent evaluation loop, and does execution-guided repair still help after accepted solutions are checked against hidden tests?

MacCode Lab



Hidden tests are isolated from generation. They are decoded and executed only after generation has finished.

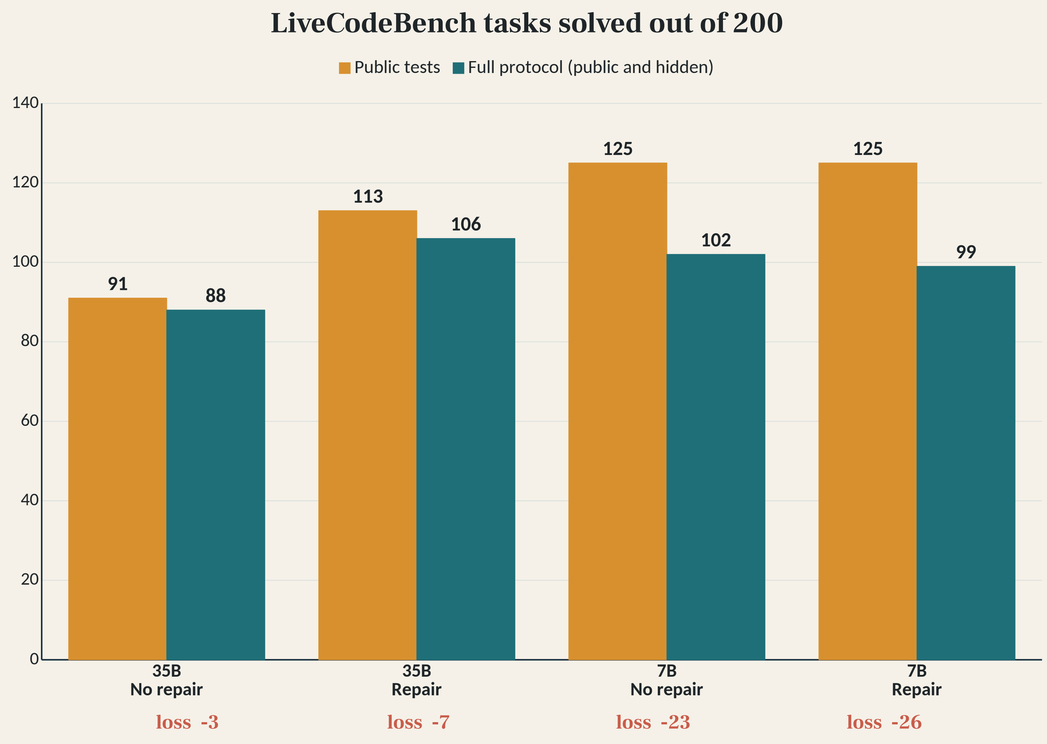
Experimental setup

Apple M5 Max, 128 GB unified memory, MLX
Qwen3.6-35B-A3B, 3-bit MLX
Qwen2.5-Coder-7B-Instruct, 4-bit MLX
LiveCodeBench v6, first 200 tasks
HumanEval, all 164 tasks
Greedy decoding and the same harness
Repair condition: 3 initial styles, 2 repair rounds, beam 2

The 35B LiveCodeBench repair run completed 200 tasks in about 2.4 hours.

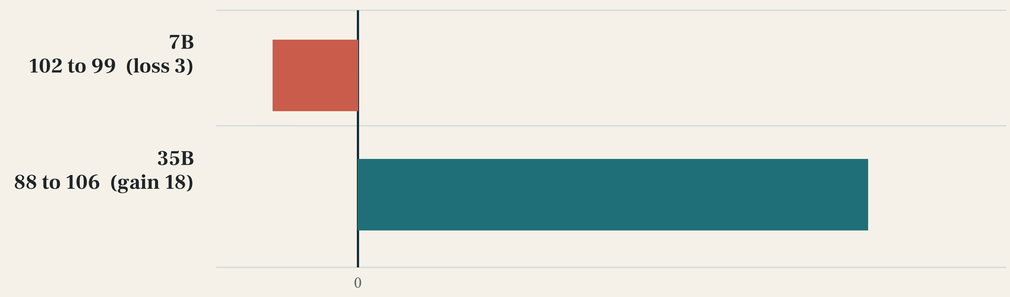
Hidden tests change the result

The two models lose very different numbers of public-test passes when accepted solutions are checked using the full protocol.



Final score difference under the full protocol

repair condition minus no-repair condition



Public scoring favors the 7B model. Full-protocol scoring narrows the gap and reverses the observed ordering in the repair condition.

With repair, the observed full-protocol score is 106 for the 35B model and 99 for the 7B model.

Wilson 95% intervals overlap. This is an observed ordering on one frozen task slice, not a statistically significant model-ranking result.

No-repair and repair are independent single runs. The repair condition also receives execution feedback and additional attempts.

When does repair help?

No repair compared with repair (independent runs)

35B model

Public tests: 91 to 113
Full protocol: 88 to 106

The repair condition scores higher in this run.

7B model

Public tests: 125 to 125
Full protocol: 102 to 99

The repair condition provides no net gain.

HumanEval was near saturation: repair changed either model's final score by at most one task.

What the prompt policy learned

35B LiveCodeBench repair run

Easy tasks	Direct and careful prompts were preferred.
Medium tasks	Structured and template prompts were preferred.
Hard tasks	No prompt style produced a strong signal.

For medium tasks, scaffolded styles reached an approximate success rate of 0.25 per attempt, compared with about 0.13 for terse styles.

Why the measurements are auditable

- ✓ Every candidate was executed before being counted.
- ✓ Hidden tests never affected prompts, repairs, rewards, or stopping.
- ✓ Per-task records and accepted solutions were persisted.
- ✓ Run status and bandit state were saved for inspection.
- ✓ Every reported count was independently re-derived.
- ✓ Accepted solutions were replayed on public and hidden tests.

Scope

Two quantized models, one workstation, a deterministic 200-task LiveCodeBench prefix, and one run per configuration. The conditions are not paired or compute-matched. Results are a reproducible systems checkpoint, not a leaderboard or a general model ranking.

Measure repair for each model, and report public scores with full-protocol re-scoring.

MacCode Lab makes a local evaluation run inspectable, resumable, and replayable.

Reproducibility materials described in the paper: harness · frozen JSONL records · cached tasks · accepted solutions · independent verification

PAPER #15

openreview.net/forum?id=nsQjG3V0Fi