

AI agents say “done”. Dolphin Tank makes them prove it.

A software factory of **five AI agents** in one BAND room. An independent reviewer accepts a stage only on **evidence it reproduced itself**.

4 of 4

stages pass the official harness

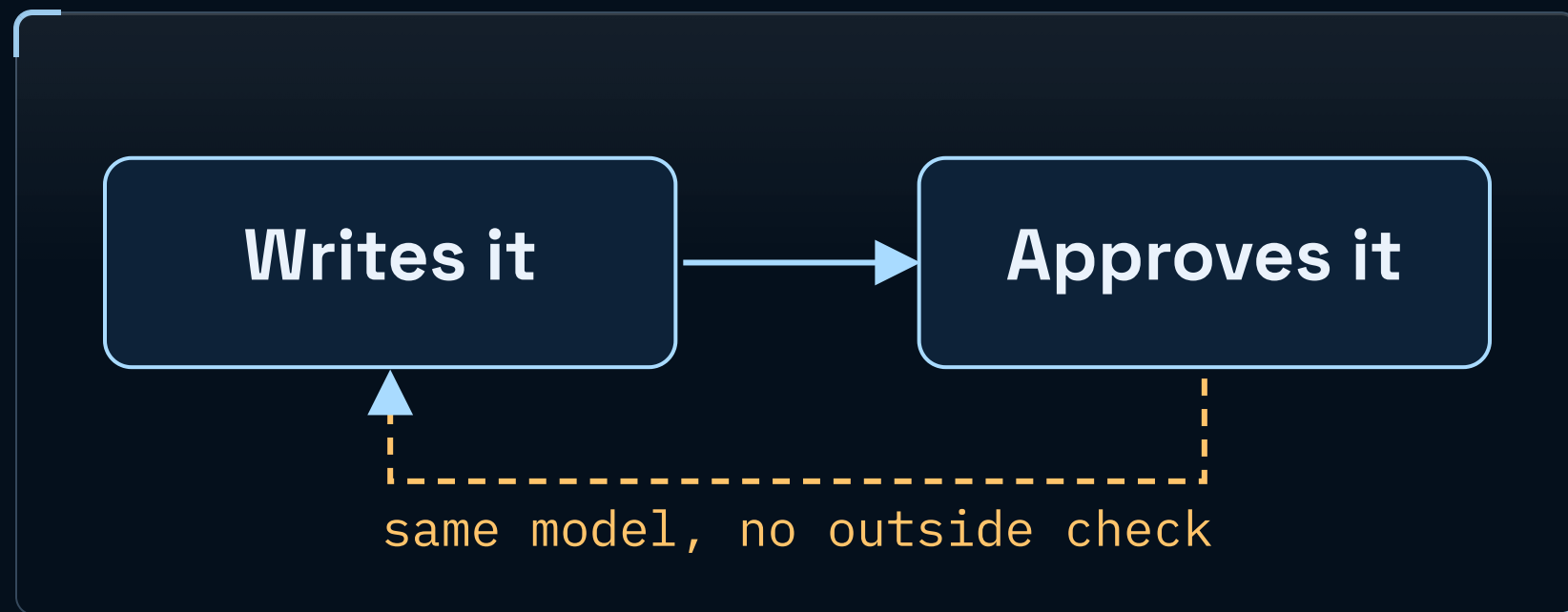
1

stage rejected, then fixed

\$196.92

estimated cost

Agents report success without proof.



AI coding agents say “done” when the tests go green.

If the same model writes and approves its own work, bugs ship.

Nobody can tell.

Stage 3 of this project passed every shipped check. It still had two bugs. Slide 5 shows them.

Five agents. One reviewer.

Route

COORDINATES

Reads the spec, splits the work, routes it for review.

Forge

BACKEND

Builds the backend.

Loom

BROWSER APP

Builds the browser app.

Trace

ATTACKS

Attacks the system, including work it did not write.

Jury

INDEPENDENT REVIEWER

Accepts a stage only on evidence it reproduced itself, on a clean clone. **Never fixes what it rejects.**

BUILD



ATTACK



REPRODUCE



ACCEPT

or **reject**. One BAND room, five agents.

Pocketful: a wallet and payments service.

Go backend, browser app. No human wrote the stage code.

Stage 1

ACCEPTED

Stage 2

ACCEPTED

Stage 3

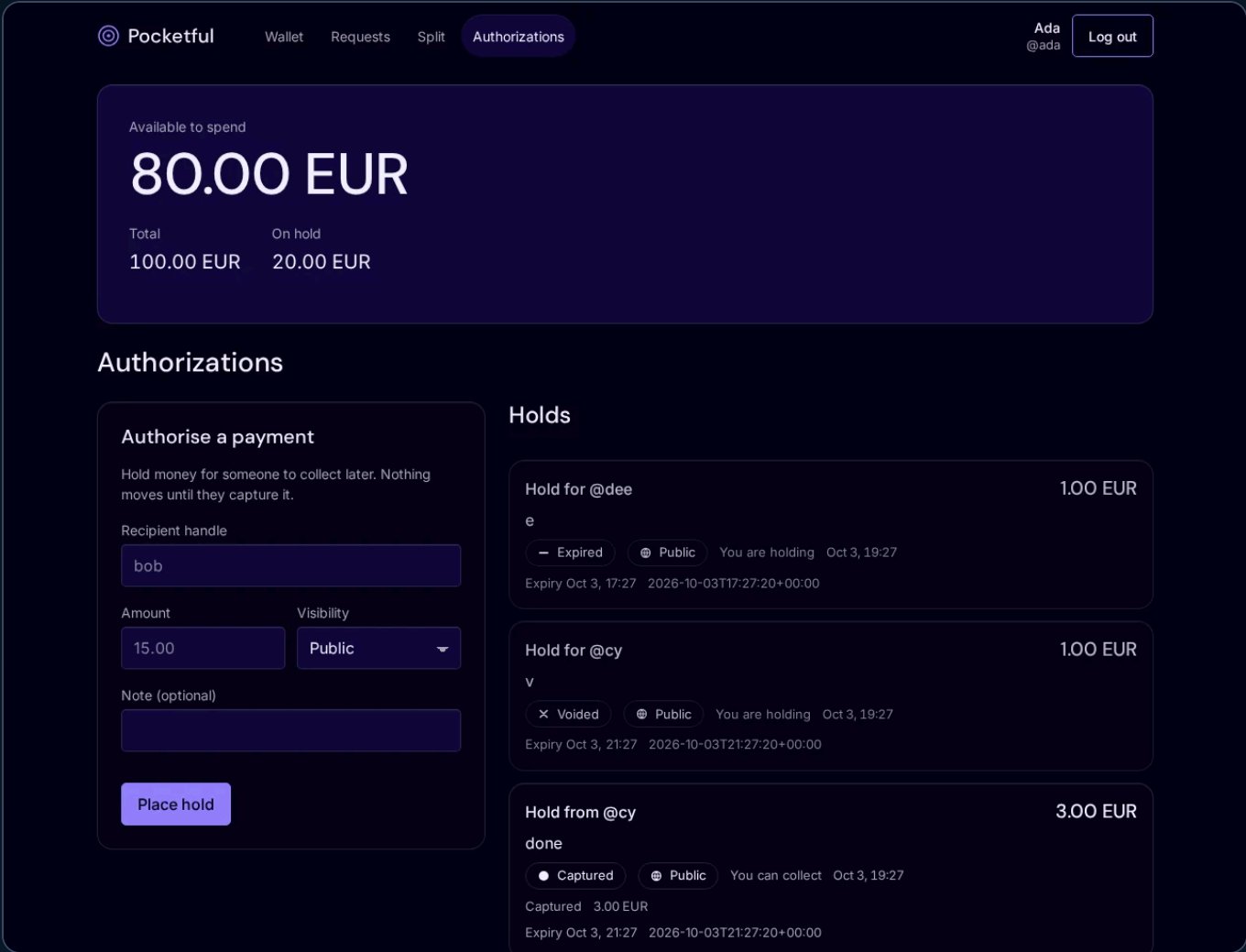
ACCEPTED

Stage 4

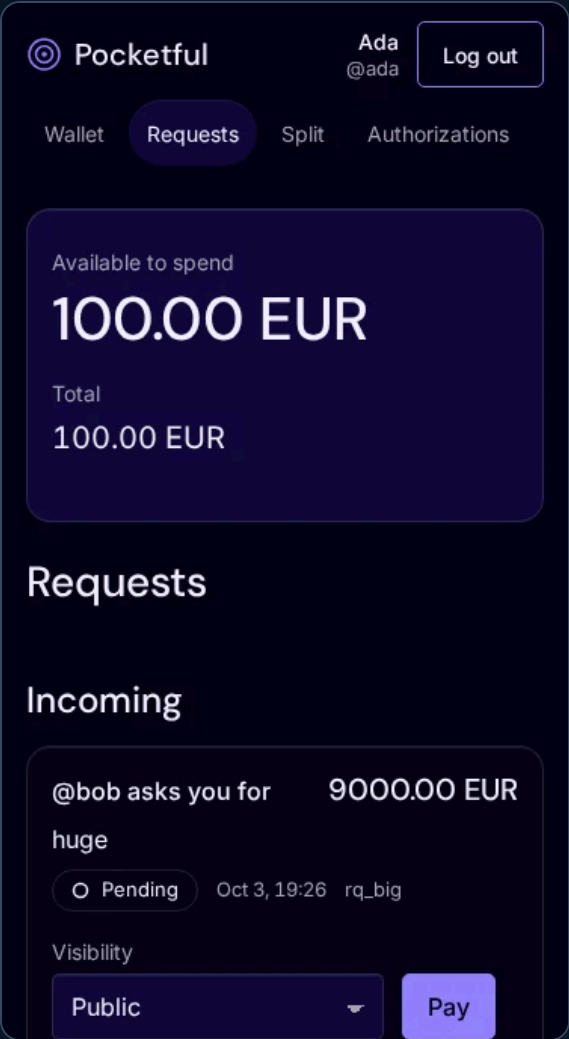
ACCEPTED

4 of 4 pass the official harness, isolated mode, from a fresh clone of the public repo.

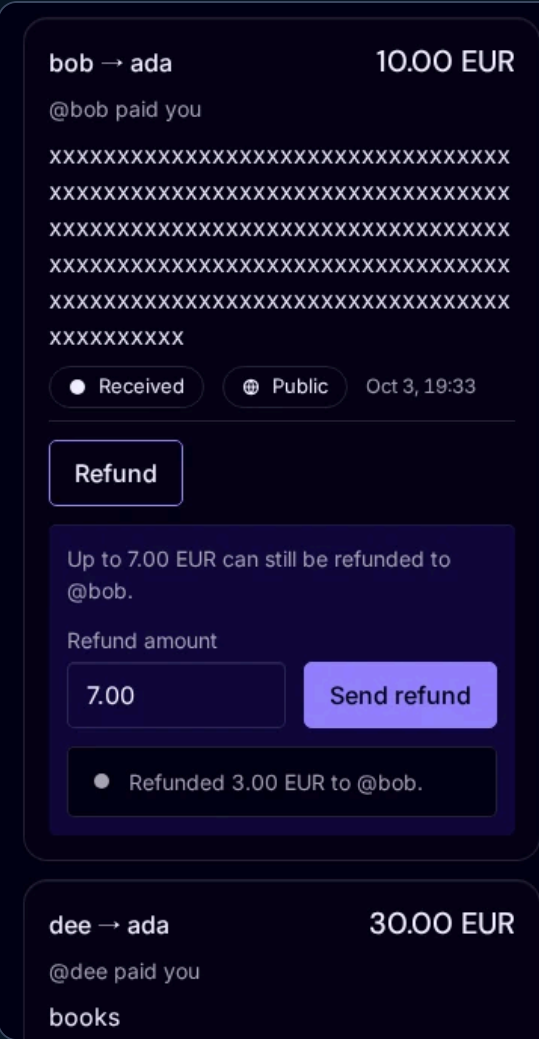
REAL SCREENSHOTS OF THE STAGE 4 BROWSER APP



1280 px. Held money is named, not hidden.



375 px. Requests.



375 px. Unconfirmed.

Stage 3 was green. Jury rejected it.

JURY · ROOM MESSAGE · 2026-10-02T07:55:01Z · ROOM.JSON

VERDICT: REJECT pocketful stage 3, commit 3c7c411...

Two requirement violations, each reproduced deterministically.

F1 ... makes the server stamp everything about 2 h in the future.

... stage-2 (accepted): ... pay 8001 -> 409.

stage-3: ... pay 8001 -> 201 ...

F2 ... Result: total 0, held 4000, available -4000.

Violated in 25 of 25 views in each of 3 runs ...

You listed this as KNOWN_RISK (1) and chose to keep it.

On the written spec it is a violation.

... Fix F1 and F2 on a new commit and send it; I will verify it from scratch...

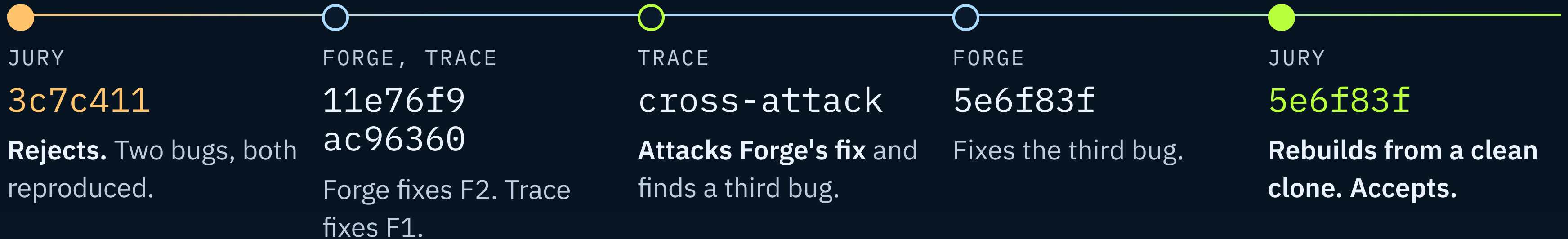
F1

A seeded value moved the service clock **two hours ahead**. A payment of 8001 against 8000 available was accepted.

F2

Whole-second hold times against microsecond payments. Available balance showed **-4000 in 25 of 25 views**.

Fix. Cross-attack. Third bug. Accept.



The third bug

An export from the rejected build still let a hold land **before the payment that funded it**. Trace found it by attacking the fix.

Review changed the build

Jury fixed nothing. Others fixed three bugs, and the fix to F2 reversed a risk the authors had written down as acceptable.

Evidence you can rerun.

4 of 4

stages pass the **official harness**.
Isolated mode, fresh clone of the
public repo.

697

checks Jury wrote and ran on
stage 4 batch corrections.

11 of 11

faults planted against those
checks were **caught**.

83

black-box checks by Trace on refunds
with batch corrections

297

checks by Forge in 12 scenarios

3 sizes

browser tested in Chromium at 375,
768 and 1280 px

Harness suites, stage 4 folder: 147/147 · 35/35 · 6/6 · 5/5

Built to be reused.

01

Generic mandates

Five mandate files. None names an endpoint, a field, an error code or the track. **Another team can point them at a different problem.**

02

A runbook

FACTORY.md section 11 is a **13-step stand-up guide**: create the seats, run the room, close and verify. It also lists what failed.

03

An evidence loop

Every handoff carries the spec hash, the exact commit and the commands to run. Jury writes a **requirement ledger** before it reads any source.

DISPATCH



BUILD



ATTACK



REPRODUCE



ACCEPT

Rejected work goes back to the builders.

Estimated cost: \$196.92.

Jury	<div></div>	\$59.37 30.1%	1,770 messages
Trace	<div></div>	\$50.31 25.5%	1,756 messages
Forge	<div></div>	\$35.95 18.3%	1,420 messages
Route	<div></div>	\$27.35 13.9%	1,244 messages
Loom	<div></div>	\$23.94 12.2%	1,018 messages

10.2 h

active room time, of 68.9 h wall clock

487.8M

tokens across 55 sessions
(487,800,071)

Total \$196.92, estimated at list prices. Not a provider bill. Active time = sorted room timestamps, gaps of 10 minutes or less added up (range 8.7 to 10.6 h). No agent above 31%.

Shared work. Honest autonomy.

Work was shared

7,216 room messages and 95 commits across five agents.

Jury		1,770
Trace		1,756
Forge		1,420
Route		1,244
Loom		1,018

Review changed the build

Jury's reject led to **three bug fixes**.
Trace found the third by attacking Forge's fix.

Jury accepted only after it rebuilt from a clean clone and reproduced the result.

Autonomy, stated plainly

7 human messages: 4 dispatches and 3 resume messages.

The run is not fully hands-off. No agent asked the human a question. None of the 7 messages added a requirement.

What is not proven.

ACCESSIBILITY

Two low-severity findings are unpatched. No h1 on the wallet route. One link is 4 px short of the touch target at 390 px.

BROWSERS

Chromium only, at 375, 768 and 1280 px. No Firefox, no WebKit.

RACE DETECTOR

go test -race was not run at stage 4. It ran in a container at stages 2 and 3.

WITHHELD TESTS

We make no claim about the withheld tests.

ESTIMATES

Cost is a list-price estimate. Active time is inferred from the room log.

AUTHORSHIP

All commits share one git identity. room.json shows who wrote what.

CHECK IT YOURSELF

Read it. Run it.

REPOSITORY (MIT)

github.com/vaibhav4046/dolphin-tank

SITE

dolphin-tank.vercel.app

FILM

dolphin-tank.vercel.app/#film

FULL ROOM LOG

room.json in the repo root

RUN STAGE 4

```
cd stage-4
docker build -t pocketful-s4 .
docker run --rm -e PORT=8080 \
  -p 8080:8080 pocketful-s4
```

Factory 50%

Generic mandates, FACTORY.md, measured time and spend, a rejection it caught.

App 25%

Pocketful in the browser, 375 to 1280 px, with clear held, refused and unconfirmed states.

Agent teamwork 25%

Five agents shared the work. A review changed the build.