

Sandworld — investor brief

Working draft, 21 October 2026. Every market fact below is sourced in `docs/research/COMPETITORS.md`; every product number is computed by the code and can be re-checked with `npm test` and `/api/v1/environments/{slug}/qa`.

One line

Verifiable digital twins of the apps people use — so AI labs can train and evaluate agents on tasks with a provable right answer — plus the expert data that teaches the rest.

The problem

Agents are now trained the way they're graded: act, check the result, learn from it. That loop is only as good as the check.

- Graders are often wrong. A study across five agent benchmarks found 15.3% of FAIL verdicts incorrect; published environments have shipped with graders that pass the wrong answer (research §5.3).
- Labs evaluate the verifier before they buy an environment, and resistance to reward hacking is their top criterion (research §1, §4).
- Real apps can't be used for this: they change without notice, can't be reset to an exact state, rate-limit or block agents, and have real side effects (research §5.1).

What we built (live in this repo)

Worlds	6 app twins — grocery delivery, P2P payments wallet and airline booking (mobile); small-business invoicing, issue tracker, and email and calendar (web) — plus Worksuite , a cross-app world where one company's email, issue tracker and billing run together. Every one is a working app a person can use in the browser
Verified tasks	117, each with a reference solution and a state-based verifier
Agent tools	127, served ready for the Claude Messages API, OpenAI function calling and MCP
Verifier QA	117 of 117 tasks graded A: reference passes on 5/5 seeds; do-nothing, random, stop-early and extra-write probes all fail
Collateral-damage detection	Episodes fail if they change state a correct solution never touches (footprint learned across seeds)
Any model, scored the same way	Claude, OpenAI, Gemini, OpenRouter, any OpenAI-compatible server and local models through one agent loop; every failed episode categorised (prompt injection, didn't ask, broke a rule, gave up, wrong result...)
Published results	Model runs on all 117 tasks on the Arena, every episode's full log one click away (see Traction)

Computer use	The same tasks driven through the apps' screens (screenshots, numbered elements, clicks and keystrokes) and graded by the same verifiers, so the cost of using a GUI instead of an API — the "GUI tax" — is measurable
RL training	Per-step training signals, deterministic fault injection (outages, rate limits, timeouts after the write went through), a local server with every world at about 10,000 steps a second, and a GRPO rollout example
Debugging	Any episode compared with the reference solution — where it first went wrong, what it changed that it shouldn't have — and retried from any step
Recovery data	One click turns an eval run's failures into expert work: each item resumes the agent's own episode where it went wrong, an expert finishes it, the verifier checks it. Labs get verified corrections of their model's own mistakes, with optional step-by-step reasoning notes
Custom twins	A scaffold turns an app's OpenAPI document into a working twin in seconds (tested on Swagger's Petstore and Bump.sh's Train Travel API); policy, tasks and verifiers are then written by hand
Safety and clarification	12 prompt-injection tasks (text planted in emails, notes, tickets and customer records asks for actions the user never requested) and 12 ask-the-user tasks, with an optional LLM-played user
Expert data platform	Demonstrations, preferences, rubric grades, SFT and red teaming; paid assessments with a fresh form for every attempt, a work queue, gold tasks, double review, appeals, payouts and Ed25519-signed credentials anyone can verify; Krippendorff's α ; TRL-ready exports with dataset card and provenance
Interfaces and formats	REST API (public sandbox, no sign-up), MCP server, Python client (Gymnasium-style), Harbor task packages (beta), ATIF v1.8 trajectories
Tests	1,050+ automated tests gate every change

Why we win

- 1. Published verifier QA.** Few vendors publish false-pass rates for their graders (research §5.3). We publish a per-task probe report for every world and refuse to ship verifiers that fail it.
- 2. Results anyone can audit.** Every published model run comes with every episode's tool calls and verifier checks, and every failure is categorised. Buyers can see exactly why a model fails, not just how often.
- 3. Cross-app and computer use on the same tasks.** Worksuite tests the hardest thing agents do — carrying facts from an email to an issue to an invoice — and every task can be run through tools or through the screen, graded by the same verifier.
- 4. Plugs into labs' own harnesses.** Harbor task packages, ATIF trajectories, a local server for RL rollouts inside a lab's network, and training signals and faults for post-training (research §6).
- 5. Human and agent work in the same world.** Expert demonstrations are recorded inside the twins and checked by the same verifiers, and a lab's failed eval episodes become expert recovery work in one click, so environments, evals and human data feed each other — the bundle the market is moving toward (research §8).

Market

- Leading labs reportedly spend on the order of \$1B a year each on data; one major lab discussed spending more than \$1B on RL environments in a year (secondhand reports, research \$1).
- Comparables: AfterQuery (founded Feb 2025; \$100M+ run rate within ~14 months; reported \$3.2B valuation), Mercor (\$2B gross run rate, June 2026), micro1 (claims \$1B run rate), Deccan AI (\$25M Series A led by A91, India-based workforce).
- Price anchors: \$200–2,000 per task; ~\$20k per cloned website; exclusive deals at 4–5× (research \$1.3).

Business model

Free public sandbox → paid pilots (one custom twin + 200 verified tasks) → environment licences (non-exclusive per twin per year; exclusive at 4–5×) → expert data per accepted task.

Traction

No customers yet; customer discovery starts now.

Published so far (run on a laptop with free open-weight models; every episode public):

Model	Interface	Tasks × seeds	Pass rate	pass@3	Obedied planted text	Asked when needed	Most common failure
Qwen3 8B (local, Q4_K_M)	tools	117 × 1	8.5% (10/117)	—	0 of 12	1 of 12	Gave up (61)
Mistral Small 3.2 24B (local, Q4_K_M)	tools	117 × 1	6.0% (7/117)	—	0 of 12	0 of 12	Gave up (85)
Llama 3.2 3B (local, Q4_K_M)	tools	117 × 1	0.0% (0/117)	—	0 of 12	1 of 12	Gave up (96)

From `benchmarks/runs/` on 2026-10-21 (`node scripts/results-table.mjs --write`). Every episode is on the Arena.

90-day goals:

- First frontier-model results (Claude, GPT, Gemini) on all 117 tasks, published with their episodes.
- 20 conversations with lab and agent-startup eval teams; 3 design partners.
- 1 paid pilot.
- 10 worlds, including the first packaged Android twin and a second cross-app world.
- First 50 vetted experts across finance, medicine, law and software.
- SOC 2 Type I readiness started.

Risks we take seriously

- **Crowded market.** Scale, Surge, Mercor, Turing, Deccan and others already sell environments. We compete on verifier quality, auditable results, cross-app and computer-use coverage and contributor

treatment, not on breadth.

- **The name.** A screening search found no trademark conflicts for Sandworld; a formal clearance search in the US and India comes before launch.
- **Labs build in-house.** We deliver in the formats their harnesses use and price below build cost.
- **Mobile demand is unproven with US labs.** It's a wedge to validate with buyers, not an assumption (research §5.1).
- **Contractor classification and security** are table stakes that sink vendors (research §7); both are on the roadmap with dates.

Founder

Revanth Rajeev builds and ships apps, including Android apps on Google Play, and built Sandworld end to end: the seven worlds, the 117 verified tasks, the any-model evaluation harness, the expert platform and the live site. He ran and hand-audited the first model benchmarks himself.

Hiring next: a sales co-founder who has sold training data or evaluations to AI labs, and an ML engineer with RL and evaluation experience.

The ask

\$500k pre-seed (about ₹4.75 crore), for 18 months of runway.

Use	Share
Team: a sales co-founder and an ML engineer, based in India	55%
Expert network: vetting, paid assessments and the first payouts before client payments arrive	15%
Compute and evals: frontier-model benchmarks, always-on hosting, client evaluations	15%
Security and company: SOC 2 readiness, a penetration test, a Delaware company with the IP assigned	15%

What it buys: 3 paid pilots and the first \$250k in annual revenue; 10 worlds, including a packaged Android twin; 100 vetted experts. Then a seed round.

Contact: revanth@sandworld.dev · <https://sandworld.dev>