

Kettleford Labs compared: agent evaluation and environment providers

Kettleford Labs, 28 October 2026

In short

Money is flowing into agent evaluation and training environments. Arena reached \$100M in annualised evaluation revenue about eight months after launch. Environment startups such as Halluminate, Fleet and Bespoke Labs have raised \$30–50M rounds this year. Labs pay six to seven figures a quarter for environment contracts. Almost none of these providers publishes prices.

Kettleford Labs fills a gap in that market. The open research simulations (Microsoft, Google DeepMind, Sierra) are free but not sold or supported. The funded vendors mostly build replicas of work: coding, browsers, finance and office apps. The closest idea to ours, Andon Labs' Vending-Bench, runs one business per agent and is not for sale. Kettleford is a whole town economy with many agents, exact replay and conduct checks, playable through structured actions, tools, a browser or code, and sold with published prices.

The providers compared

Figures are the latest public ones as of 28 October 2026. "Not public" means we found no published price or figure. "Reported" marks figures that come only from secondary sources.

Provider	What they offer	Pricing	Funding raised	Revenue
Kettleford Labs	Long-horizon, many-agent town economy for evaluation and training: 23 scenarios, four interfaces (API, tools, browser, code), safety and oversight pack, payment-agent tests, conduct reports, collusion studies, decision audits, training packs, Inspect and Environments Hub packages, public Index	Studies \$5K–\$25K; subscriptions \$1.5K–\$5K a month; custom scenarios \$20K–\$100K; lab licences \$100K–\$500K a quarter	Self-funded	Pre-revenue
Andon Labs	Vending-Bench 2 (one business for a simulated year), Vending-Bench Arena (a few competing agents), real AI-run shops	Not public; run in-house for labs	Y Combinator (W24); amount not public	Not public
Arena (formerly LMArena)	Community-vote model evaluations with audit samples and SLAs; free public leaderboard	Not public ("contact us")	About \$450M: \$150M Series A (Jan 2026, \$1.7B); \$200M Series B (Oct 2026, \$3.1B)	\$100M annualised (Jun 2026)
Prime Intellect	Open Environments Hub, open RL training stack, hosted training	Hub free; hosted training by GPU-hour	Over \$150M; \$130M Series A (Jul 2026, about \$1B)	Over \$100M annualised (Jul 2026)
Mechanize	Long-horizon software-engineering	Not public	\$9.1M at \$500M (Apr 2026); Google talent and	Not public

Provider	What they offer	Pricing	Funding raised	Revenue
	environments for labs		licence deal reported at over \$1.5B (Sep 2026)	
Halluminate	Finance training environments and benchmarks	Not public	\$38.5M; \$30M Series A (Oct 2026)	"Mid-eight figures", profitable (self-reported)
Fleet	Environments for browser, coding and chat agents	Not public	\$45–50M Series A at about \$725–750M (reported, 2026)	About \$60M annualised (reported, Apr 2026)
Bespoke Labs	Containerised, verifiable environments and expert data	Not public	About \$40M seed and Series A (Jul 2026)	Not public
Deeptune	Workplace "training gyms" (Slack, Salesforce, Excel)	Not public	\$43M Series A (Mar 2026); bought by Mercor (Jul 2026)	Not public
Surge AI	Expert human data, evaluations and off-the-shelf RL environments	Not public	Bootstrapped; talks at \$15–25B reported (2025)	About \$1.2B (2024)
Mercor	Expert marketplace for post-training; RL environments via Deeptune	Not public	About \$484M; \$350M Series C at \$10B (Oct 2025)	\$2B gross annualised (Jun 2026)
Scale AI	Data, evaluations (SEAL leaderboards), managed RL environments	Not public	Meta paid \$14.3B for 49% at \$29B (Jun 2025)	About \$870M (2024)

Provider	What they offer	Pricing	Funding raised	Revenue
nof1 (Alpha Arena)	Models trading real money live, as a public benchmark	Free to watch	\$15M (May 2026)	Not public
METR	Non-profit dangerous-capability and autonomy evaluations	Not sold	About \$71M in philanthropy (2026); takes no money from AI companies	Non-profit
Magentic Marketplace (Microsoft)	Simulated two-sided market of buyer and seller agents	Free, MIT licence	Research project	None
Concordia (Google DeepMind)	Social simulation library with a game master	Free, Apache-2.0	Research project	None
tau-bench (Sierra)	Customer-service agent simulation	Free, MIT licence	Sierra raised \$950M at over \$15B (May 2026)	Sierra about \$150M annualised (Feb 2026)
GDPval (OpenAI)	1,320 real work tasks across 44 occupations; 220 public	Gold set free; no licence stated	Research project	None

Revenue figures are not like for like. Mercor's is gross, before contractor pay of 60–70%, and some startups annualise a single quarter.

What the market pays

Few providers publish prices, but interviews and reporting give a consistent picture of what labs pay.

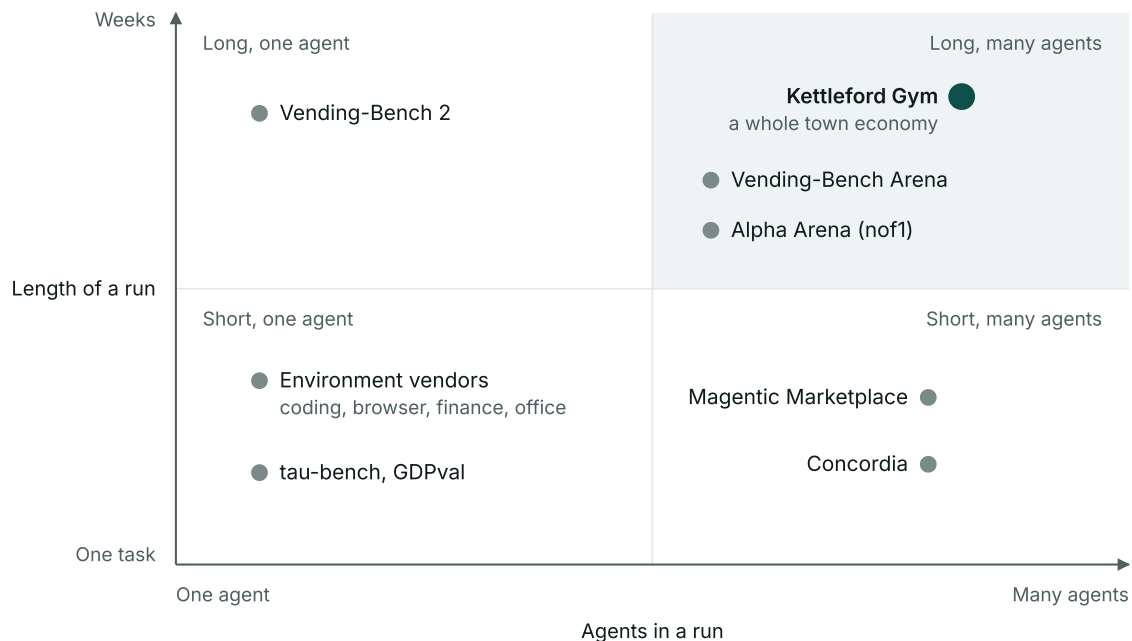
Measure	Typical figure	Source
Price per RL training task	\$200–\$2,000; about \$20,000 for rare, complex software tasks	Epoch AI , Jan 2026
Environment contracts with labs	Six to seven figures a quarter; one lab saw \$300K–\$500K	Epoch AI , Jan 2026
Premium for exclusive environments	About 4–5 times the non-exclusive price	Epoch AI , Jan 2026
Cost to build an environment	About \$20K for a website replica; about \$300K for a high-fidelity app such as Slack	Epoch AI , Jan 2026
Suggested spend per task	At least \$500, rising to a few thousand dollars	Mechanize , Aug 2025
Lab budgets	Anthropic discussed spending over \$1B on RL environments in a year (a discussion, not a confirmed spend)	Epoch AI FAQ , Sep 2025
Training compared with benchmarks	Labs want training environments about 10–20 times as often as benchmarks	Epoch AI , Jan 2026

Kettleford's lab licence (\$100K–\$500K a quarter) sits inside the reported contract range. Its study price is below a single high-end environment build. Paid evaluation can also grow quickly on its own: Arena went from launch to \$100M annualised in about eight months.

Where Kettleford differs

Only Kettleford runs many agents over weeks in a whole economy

Where each offering sits by length of a run and number of agents; positions are zones, not measurements



Positioning by length of a run and number of agents, nine offerings.

Most paid environments test one agent on one task. The arenas add a few competitors, and the open market simulations add many agents but run short episodes. Kettleford combines all three: many agents, weeks of town life and a full economy.

- **What we offer that others don't sell:** exact replay, fast forks and saved states from any moment; the same task through four interfaces, scored by the same ledger; conduct and oversight tests (price-fixing, theft, deception, phishing, hiding losses from an auditor, sandbagging, behaving differently when unwatched) with quoted evidence; population tests with dozens of copies of one model; tests for agents that shop and pay for people; graded conduct reports and collusion studies; adversaries that adapt to the agent; and published prices.
- **Where others are ahead:** the funded vendors replicate real enterprise software and coding at a breadth our nine in-town apps don't attempt, and that is where most lab spending goes. Arena and the vendors also have track records with labs that we have yet to build.
- **How we fit alongside them:** a judgement and conduct test that complements work suites, with open scenarios and a public Index to earn trust.

Sources

All figures are linked in the tables above and were checked on 28 October 2026. Funding and revenue for Fleet and Surge's valuation talks come from secondary reports. Arena's \$100M figure is self-reported and dates from June 2026; its October Series B repeated it. Andon Labs' funding beyond Y Combinator is not public, and aggregators disagree on it.