Engineers who train and evaluate AI models.

Awana staffs working senior software engineers, in dedicated pods, for human data, coding benchmarks and red-teaming. They work inside your tools, under your access rules, in US time zones.

What a pod does.

Human data

Preference and RLHF data with written rationales, rubric grading, and review of agent and computer-use trajectories.

Coding benchmarks

Repo-level tasks written, solved and graded: a real codebase, a failing test, a reference fix.

Red teaming and evaluation

Attack sets mapped to the OWASP Top 10 for LLMs, findings rated by severity, and a retest once you have fixed them.

How it starts.

Free

Calibration round

A blind test against your gold set, before you pay for anything.

Week 1

Paid trial

Two to three engineers on your task type, inside your platform. One fixed price for the week.

End of the week

Read-out

Acceptance rate, rationale samples and agreement. A sample is on the next page.

Your call

Keep or walk

No obligation. Run us beside your current vendor on the same tasks and compare.

Your environment, your rules.

You control

- Your platform, tools and virtual desktops
- Your SSO, least privilege, logged sessions
- Your access tiers and redaction rules
- Your tasks, rubrics and data, which stay in your environment

Awana guarantees

- NDA and a standard DPA before any access
- A background check on every evaluator
- No datasets of our own: we do not build or sell training data
- One accountable lead for each pod

The record.

50	1 in 40	~3 weeks	3	~Half
engineers on AI-training work for a frontier AI lab in three months, name not published	engineers make it through our vetting	to staff most pods, calibration included	working languages: English, Spanish, Portuguese	of US rates for the same seniority

Awana is United States-led, and its leadership holds an active US security clearance. Legal entity: Hire Llama Inc., Florida.

Scope your first pod. The calibration round is free.

Thirty minutes on the task, the rubric and your platform. letsgrow@awana.io · awana.io/ai-training

Book a call

SAMPLE. This page shows the format of the read-out a lab receives after a one-week trial. Every number, task and quote on it is invented. It does not report results from any client.

Trial read-out: coding-benchmark pod, week 1

Pod	Task type	Rubric	Agreement floor
Three engineers	Repo-level bug fixes, Python and TypeScript	Yours, version 1.2	80%, set by you

The week in five numbers.

42	36	5	1	86%
tasks shipped	accepted on first review (86%)	accepted after one revision	rejected	agreement with your gold set, against a floor of 80%

By engineer.

Engineer	Tasks	Accepted first time	Agreement with gold set	Our read
A	16	15 (94%)	91%	Keep. Strongest rationales in the pod.
B	15	13 (87%)	88%	Keep.
C	11	8 (73%)	76%	Below your floor. We swap this seat within the week.

Two rationales, as written.

> "The reference fix misses the empty-list case. I added a failing test for it first, then the fix. All hidden tests pass."
> Task 17 · accepted on first review

> "The patch passes the visible tests by special-casing the fixture path. The hidden test for a second path fails, so this is a fail against criterion 2."
> Task 31 · graded as a fail, with a minimal repro attached

What we flagged in your rubric.

Four items graders read two ways. Example: criterion 3 does not say whether a refactor outside the diff counts against the task. Each is logged with the task it came from, for your researchers to rule on.

Our recommendation.

Keep A and B and replace C. Add four engineers in week 2 and three in week 3 to reach ten, each calibrated against your gold set before they start. If the numbers above do not beat your current vendor on the same tasks, walk.