

Series B+ · Montreal, Canada

Lanternfish Labs

Research lab building methods to oversee and verify AI systems more capable than the people supervising them



Lanternfish Labs

What's broken today

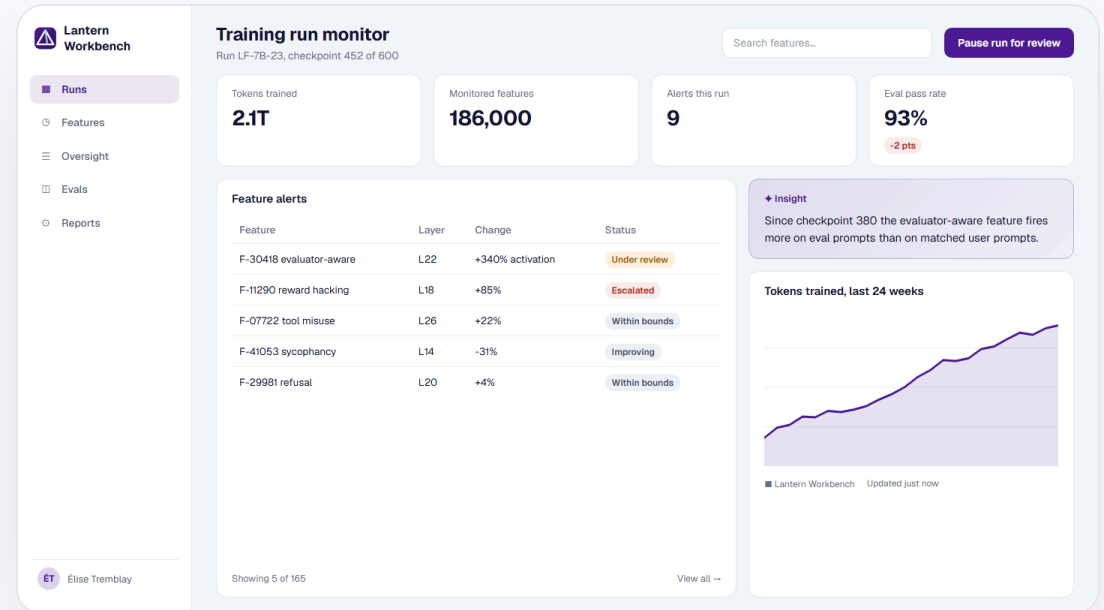
As models get better at long, multi-step tasks, human reviewers can no longer check every step, and most evaluations test what a model says rather than what it is doing internally. Developers already ship systems whose failure modes are found by users weeks after release. There is no reliable, shared method for confirming that a much more capable system is pursuing the goal it was given.



Our solution

Lantern Workbench

Lanternfish trains its own research models up to mid-frontier scale so it can study oversight with full access to weights and training data. Its work covers interpretability (mapping internal features to behaviours), scalable oversight protocols in which weaker models and humans check stronger ones, and pre-deployment evaluations for deception and goal drift. Methods and benchmarks are published; tooling is licensed to other developers.



Lanternfish Labs

Oversight protocol trial

Lantern Workbench

Runs

Features

Oversight

Evals

Reports

Oversight protocol trial

Debate protocol v3, 1,200 coding tasks

Compare with v2 protocol

Best judge accuracy

87%

+11 pts

Flawed solutions caught

418/472

Human hours

64

Results by judge

Judge	Protocol	Accuracy	Cost per task
Small model	Direct check	61%	US\$0.02
Small model	Debate v3	79%	US\$0.11
Human expert	Direct check	74%	US\$14.00
Human + small model	Debate v3	87%	US\$3.20
Human + small model	Consultancy	81%	US\$2.70

Showing 5 of 266

View all →

✦ Insight

Debate helps most on tasks with hidden test failures and adds little on style or naming flaws.

Lanternfish Labs

ÉT

Élise Tremblay

Why now

The timing

Models now complete multi-hour tasks on their own. Oversight methods need to work before systems outpace the people checking them, not after.

Market

Frontier model developers, public AI evaluation bodies and enterprises deploying autonomous agents; spending on safety evaluation and assurance is small today but grows with each new requirement for pre-deployment testing.



Lanternfish Labs

How we make money

Multi-year research and tooling licences for model developers (US\$2M–US\$15M a year), paid evaluation engagements before major releases, and grant-funded open research.



Traction

Where we are

84

Researchers

37

Papers published

5

Licensing partners

11,000

GPUs

Compute secured

- 37 papers; two oversight benchmarks adopted by other labs
- 5 licensing partners for interpretability and evaluation tooling
- Pre-deployment evaluations delivered for 9 model releases



Lanternfish Labs

Team

Who's building it

ÉT

Élise Tremblay

CEO and co-founder

SA

Samuel Adjei

Chief Scientist and co-founder

MZ

Mei-Lin Zhou

Head of Interpretability

RM

Rohan Mehta

Head of Evaluations



Lanternfish Labs

The ask

What we're raising

Raising US\$450M Series C to fund research models and oversight work through 2029

- Compute for mid-frontier research models (about 60%)
- Grow interpretability and evaluation teams to 140 people
- Open benchmarks and grants for external researchers



Lanternfish Labs