Xplore.

A platform for training and running agentic AI systems.

FIRST MULTI-YEAR CONTRACT

£1M+

Total contract value. Signed.
Medical device company
cardiac screening

WHAT WE DO

We reduce the cost of building agentic systems by **10×** and make them production reliable

£144k platform year vs £400k–£1.6M manual.
66% → 97% on held-out benchmark.

AI moved to agents. Training didn't follow.

NEVER SHIP

95%

of enterprise agentic systems never reach production. They work in demos - they break on real data and real edge cases.

MIT NANDA, 2025

NO EVALUATION

48%

of agent teams ship with no evaluation at all. 74% rely on manual human review as their only quality gate.

LangChain State of Agent Engineering 2026, n=1,340 · Chanl 2026, n=306

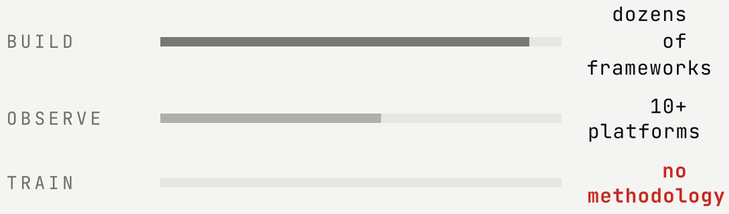
PROJECTS CANCELLED

>40%

of agentic-AI initiatives will be abandoned by 2027.

Gartner 2025

The techniques that improve a base model do not improve an agent.
Existing tools optimise only the system prompt -
- a shallow surface that cannot make an agent reliable.



LangChain · CrewAI · AutoGen build agents. LangSmith · Datadog · Arize observe. DSPy · TextGrad tweak the prompt. Nobody trains the agent.

We built the training loop for agents.

Automated training for the agent itself — not just its prompt. Real results vs. manual engineering.

Cheaper than manual engineering

↓ COST DOWN

~220×

£1.50 per iteration vs. £320–640 manually.

30 tasks × £0.05. Manual: 4 hours of senior engineer time per cycle.

Faster, no engineer in the loop

↓ TIME DOWN

~60×

4 minutes automated vs. ~4 hours manual.

30 iterations = 2 hours on Xplore vs. 120+ engineer-hours.

Reliability after training

↑ ACCURACY UP

97%

66% → 97% success on clinical trial benchmark.

30 iterations, no human intervention. 30 tasks, 6 drugs, 100 patients. Slide 12.

One training run on Xplore: **£43, 2 hours**. Same depth manually: **£10k+ and 3 weeks** of engineer time.

The industry learned to train models. We make the shift to training agents.

01 Simulate

The client's business, rebuilt as a simulation. The agent does its actual job inside - end to end - before it touches production.

02 Benchmark

Define real tasks inside the simulation. Multi-axis automated scoring - no manual review.

03 Train

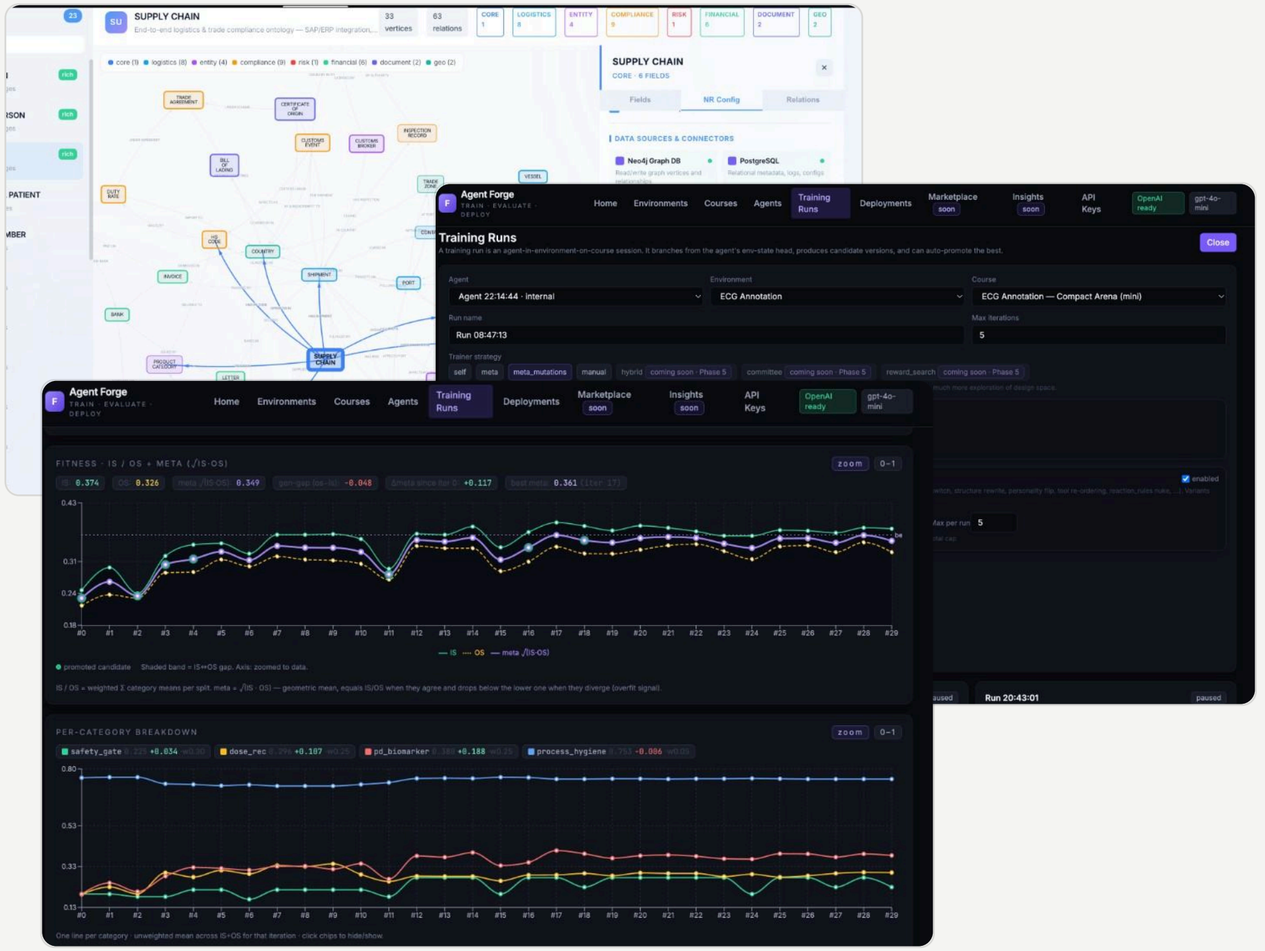
The platform improves the agent's full configuration. Each iteration scored against the benchmark.

04 Deploy

Push the trained version to production. Live data feeds back into simulations.

IN Data · rules · tools · environments · any LLM

OUT Production agent · API · dashboard · policy gates



We treat the agent as one object, not as a prompt.

Xplore co-trains the **whole system** - agent, sub-agents, tools, ontology, prompts, instructions, policies - over a single graph that is both the runtime and the training context.

PATENT · SEP 2025

UK patent priority filing on Node Resolution - graph-centred multi-agent orchestration. Provenance, audit trails, deterministic replay.

We score the result, the process, and the cost.

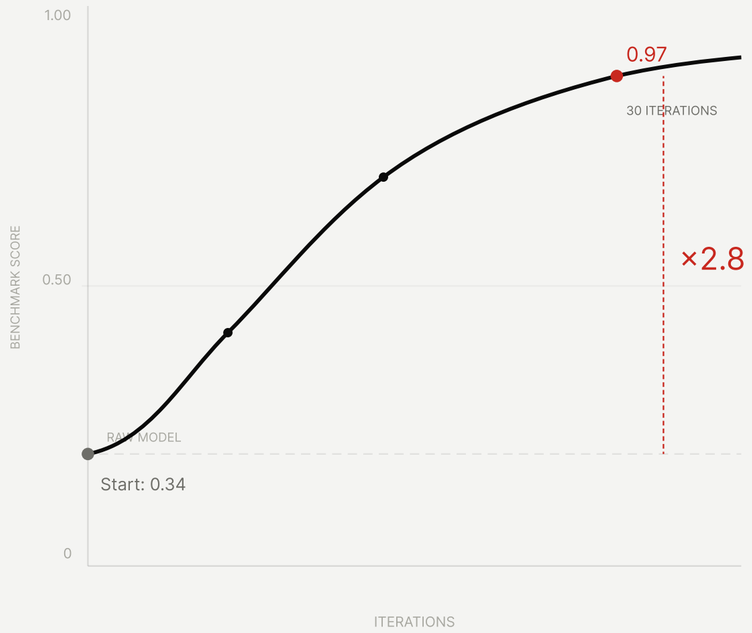
Evaluators compose into **weighted chains**. Statistical checks run first. LLM judges handle ambiguity. Every score normalised to [0, 1] and weighted by importance.

CASCADING	Cheap checks first, LLM judges last
SPLITS	IS / OS / Live - overfitting caught automatically
LEADERBOARD	Any agent, same evaluators, ground truth hidden

The agent gets better automatically.

The benchmark defines what “good” looks like. The platform runs **iterations** - find weaknesses, propose improvements, test, verify, promote. Each round the agent improves on real business metrics.

ITERATE	30 rounds, ~2 hours, ~£45 compute
VERIFY	Only proven improvements get promoted
VERSION	Every config is an immutable snapshot



Not just the prompt.

Most tools rewrite instructions and stop. Xplore changes **the whole agent system** - tools, procedures, rules, data access. The deeper the change, the bigger the impact.

L0-L1	Prompts & instructions - where everyone else stops
L2-L3	Workflow, tools, data access - Xplore today
L4-L5	Tests, sub-agents, architecture - our roadmap

Where the market is. Where we're going.

Every level adds a deeper kind of change the system can make on its own. Higher level = less manual work, faster adaptation. The market stops at L1.

			BUSINESS IMPACT	ROADMAP
L0	**Rewrite prompts**	Change instructions between runs. DSPy · TextGrad · GEPA	**~£640/fix · days.** Engineer rewrites prompt, tests by hand, repeats.	Shipped
L1	**Rewrite with context**	Use traces to improve prompts. LangSmith · Braintrust · Patronus	**~£320/fix · days.** Traces help, but still manual per change.	Shipped
L2	**Change the flow**	Add or remove steps, branches, validators. Now only manual	**£1.60/fix · minutes.** Workflow adapts automatically.	Shipped
L3	**Change tools**	Modify data sources and tool access. Now only manual	**£1.60/fix · minutes.** Adapts integrations, no rebuild.	Shipped
L4	**Change tests**	Evolve tasks and scoring with the agent. Now only manual QA	**Self-correcting.** Quality improves without human QA.	Q4 2026
L5	**Change the team**	Create new sub-agents, routers, memory. Only research papers	**Autonomous.** System scales itself.	2027

LangChain 2026 · n=1,340 · Chanl 2026 · n=306 · no vendor ships improvement beyond L1. Cost: slide 03 methodology.

Where we are.

PAYING CLIENT

£1M+

Total contract value. Signed.

Medical device company.
First agent system (ECG screening) in production.
More agents in training.

Base — annual platform license
Usage — per-run training + per-query inference
Revenue scales with client adoption.

Sector: medical devices · cardiac diagnostics · Holter ECG analysis

INDEPENDENT VALIDATION · EDINBURGH NAPIER UNIVERSITY

TRL 4

5,078 test cases · zero unhandled failures

Each architectural layer adds measured accuracy. Final report - May 2026.

TECHNOLOGY ECOSYSTEM

Edinburgh Napier University
Scottish Critical Technology Supercluster
Interface Scotland
Digital Catapult Supply Chain Hub
ScotlandIS

COMMERCIAL TRACTION

Agents trained and deployed for:

- **Cardio Risk Detection**
- **Supply Chain Disruption**
- **Clinical Trials**
- **Border Control**

First agent in production. More in training.

Medical device company. Cardiac screening agent - trained and deployed on Xplore, no Xplore engineers involved.

01 **Simulate.** Client's own data: 52 patients × 24hr Holter ECG. 7 cardiac event types (AF, VT, bradycardia, PVCs, pauses, SVT, ST changes). 10 analysis tools. Agent screens each recording end to end.

02 **Benchmark.** Cardiologist-annotated ground truth for every event. Scored on detection rate, classification (F1), temporal overlap (IoU), and report quality.

03 **Train.** 30 automated iterations, no engineer. gpt-4o-mini: .15 → .91 weighted score.

04 **Deploy.** Embedded in the client's Holter product via API. Delivers event annotations, risk flags, and clinical narrative per recording. 2 weeks from setup.

AGENT LEADERBOARD — ECG ANNOTATION BENCHMARK

#	AGENT	ANNOTATION	PROCESS	SCORE
1	**gpt-4o-mini** TRAINED	**.94**	**.98**	**.91**
2	gpt-4.5 raw	.31	.78	.38
3	claude-opus-4 raw	.27	.75	.34
4	gpt-4o raw	.22	.71	.29
5	claude-3.7-sonnet raw	.19	.68	.26
6	gpt-4o-mini raw	.08	.51	.15

Same 52 tasks, same tools. Only difference: 30 iterations of Xplore training. The weakest base model, after training, outperforms every frontier model run raw.

Asks.

01 · RESEARCH

Joint benchmarks with universities.

We bring the platform. Partners bring the domain. Together we create open, reusable evaluation standards.

Active: Edinburgh Napier · cardiac diagnostics. Two more slots in 2026.

02 · ENTERPRISES

Client data + our training = production-ready agent.

We co-author a benchmark on the client's domain, train a production agent on Xplore, and hand it over — deployed, governed, scalable.

Design → benchmark · **PoC** → train + validate · **Rollout** → deploy under client policy gates

03 · DEV TEAMS

Raw data + tasks = self-serve.

Client developers create benchmarks, train agents, and deploy — all through the platform. No Xplore engineers involved.

Pilot → onboard, first benchmark, first agent · **Rollout** → production, usage-based pricing

Edinburgh deep-tech team. Next quarter: **benchmarks and pilots**.

Methodology & sources.

A · INTERNAL BENCHMARKS (SLIDE 03)

clinical_trial_mini_30 — 30 tasks, 6 drugs, 100 simulated patients (deterministic seed, ≥18 months longitudinal EHR per patient). Evaluator chain: `safety_gate > dose_rec > pd_biomarker > ae_risk > pk`.

logistic_chain_v2 — supply-chain reasoning over manifests, vessel telemetry, registries and OSINT.

supply_chain_shock — shock-event detection across the logistics graph.

ecg_annotation_v1 — 52 tasks from 7-lead Holter recordings. 7 cardiac event types + 3 activity types. IoU + F1 scoring on 60 s windows. Client-created benchmark (slide 13).

B · NUMBER DERIVATION (SLIDE 03)

`£1.50 / iteration` — 30 tasks × £0.05 / iteration-task pair (our published rate). Manual equivalent: senior engineer at £80–160/hr × ~4h per cycle = £320–640.

`~4 min / iteration` — 30 tasks, sequential LLM call + evaluator chain. Manual equivalent: ~4h per cycle (review traces, rewrite, rerun, re-check). 30 iterations = 2h platform vs 120h+ manual.

`34% → 3%` — failure rate on `clinical_trial_mini_30` after 30 iterations. 30 tasks, 6 drugs, 100 simulated patients. Full run: £43 compute, ~2 hours. Other benchmarks: `logistic_chain_v2` 0.45 → 0.09; `supply_chain_shock` 0.52 → 0.08.

C · MATURITY LADDER (SLIDE 11) & THIRD-PARTY SOURCES CITED

Ladder · structure adapted from internal trainer-maturity reference. Levels describe what the loop is permitted to *edit* — prompt → trace-informed prompt → flow → tools & data → benchmark → architecture. Tool placement is editorial, defensible against public docs. Coverage bars show market availability, not adoption.

MIT NANDA · The GenAI Divide: State of AI in Business 2025 (Jul 2025). 5% production-rate. Method: 300+ public deployments + 52 enterprise interviews + 153 leader survey.

LangChain · State of Agent Engineering 2026, n=1,340. 48% ship without evals; 89% observability; quality #1 blocker (32%).

Chanl · Measuring AI Agents in the Real World, 2026. n=306. 74% rely primarily on human evaluation. 68% of agents execute ≤10 steps.

RAND · cited in AlphaCorp 2026. 80%+ of AI projects fail to deploy. Average sunk cost £120k+; restart costs 50–75% of original budget.

Gartner · press release 2025-06-25. >40% agentic-AI cancellation prediction by 2027.

TechCloudPro 2026 · enterprise multi-agent build cost band £400k–£1.6M, 8–18 months.

AgentList · State of AI Agent Development 2026 (n≥50). £38k average; £120k+ projects 16–32 weeks.