

PRE-SEED · OCTOBER 2026

Oxiedo

Neural networks that report what broke while they train, and fix it.

oxiedo.com/black-box

Replay the archived run, component by component

Rokib Al Dhin Raadh, Founder
raadxbusiness9@gmail.com

Runs fail silently. No one can say which part broke.

Builders pay for failed runs in compute. Validators inherit a finished black box.

\$170M

to train one frontier model (Llama 3.1, est.)

Every restart re-spends that compute.

~20

loss spikes in one PaLM run

Fix: rewind, skip data, retry. The cause is never named.

35+

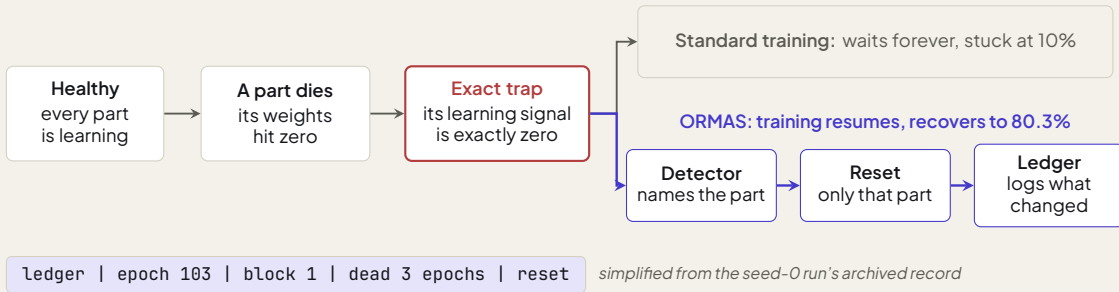
manual restarts of OPT-175B in 2 months

Divergence tracked a spike in **one layer**. Nothing named it as it happened.

\$170M: Stanford AI Index 2025 (Llama 3.1–405B). 35+: Zhang et al., OPT-175B logbook, Meta 2022 (final-layer activations).
~20: Chowdhery et al., *PaLM*, arXiv:2204.02311 (2022), §5.1 Training instability.

A dead part is a trap: we name, reset and record it.

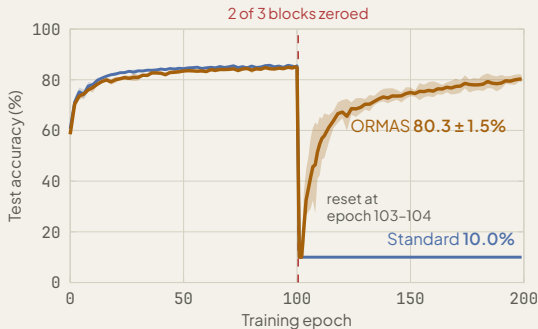
Standard training cannot leave this trap. A targeted reset can, and every reset is written down.



Trap: CIFAR-10 CNN, two of three blocks zeroed; standard runs stuck at chance in 12/12 archived runs (H-1, H-2). 80.3%: 3 seeds (H-1). Ledger entry: seed-0 run, blocks flagged at epoch 101, reset in epoch 103 (F3 D-072).

Standard stays at 10%. ORMAS recovers to 80%.

Two of three blocks of a trained image model zeroed at epoch 100. Three seeds per curve.



WHAT HAPPENED

Standard: 10.0% on every seed. Its learning signal is exactly zero.

ORMAS names the dead blocks, resets them, recovers to 80.3%.

Every step is on the record: 85 logged interventions per run, each with block, epoch and reason.

Scope: CIFAR-10, one small CNN; curves are 3-seed means, band = seed min-max. Head-to-head vs other repair methods: A2.
Curves: archived runs `ablation_{standard,ormas}_cnn_dead_s0-2`, Jul 2026 (D-070/071).

Don't trust us: replay the run yourself

oxiedo.com/black-box replays an archived run. Every figure is read from the archive.

THE RECORD · written during training

EPOCH	COMPONENT	DIAGNOSIS	ACTION	MAGNITUDE
101	conv block 1	dead	flagged	none
101	conv block 2	dead	flagged	none
103	conv block 1	dead	reset	
103	conv block 2	dead	reset	

Check it in two minutes

- Each component's health, epoch by epoch.
- The ledger: component, epoch, reason, size.
- A scenario where ORMAS loses.

Narrated tour: [Play the demo](#).

Try it now: open oxiedo.com/black-box, choose the dead-layer scenario, scrub to epoch 101 (the damage), then 103 (the reset).

Replay of archived run files; nothing is simulated in the browser. Rows: run ablation_ormas_cnn_dead_s0.

First product: a health monitor plus a change record

Runs beside their training job, read-only, on their machines. Already built; next is packaging.

TRAINING HEALTH MONITOR

Health of every component, on every run.

Names what failed, and when.

CHANGERECORD

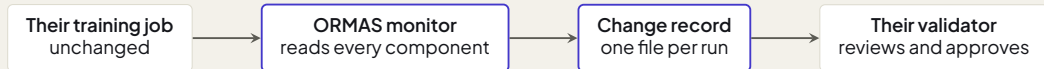
Every intervention: component, epoch, reason, size.

One file for the validator.

STATUS

Today: both run inside V1.

Next: record clean-up, then packaging.



Then, each gated by a named test: Label Audit (frozen-model test) · Source Separator (V2 separation study).

Record clean-up (V1.1): one format, exact steps, only applied changes logged, declared limits enforced. Commercial Plan 04, 06, 07.

Regulators now ask how an AI model was built

Four regulators now want evidence from training itself, not only from the finished model.

EU AI ACT · DEC 2027

Credit scoring is high-risk: document training, govern data.

Fines to €15M or 3% of turnover.

US · SR 26-2 · APR 2026

Replaces SR 11-7.

Non-generative AI models explicitly in scope, with ongoing monitoring.

UK · SS1/23

Deliberately covers AI models.

The Bank of England builds on it in 2026.

FDA · PCCP · DEC 2024

AI devices change only under a documented change plan.

1,614 AI devices listed.

Banks are not ready: large banks run 365 models; 67% use AI/ML, only 29% say their framework fits, and 75% put under 5% of staff on AI validation. Gartner: AI governance spend passes \$1B by 2030.

EU: Reg. 2024/1689 Annex III 5(b), Art. 99; dates by Reg. (EU) 2026/1744. US: SR 26-2, 17 Apr 2026. UK: PRA SS1/23; BoE letter, 1 Apr 2026. FDA: PCCP final guidance, 4 Dec 2024; AI device list, Sep 2026. Banks: Deloitte EMEA MRM survey 2025. Gartner, 17 Feb 2026.

A \$19B market by 2035, and we know who signs

Model-risk spend is already a budget line. We sell into it from training.

\$5.9B

\$149M

\$4.75M

TAM · AI model-risk mgmt, 2025;
\$19.1B by 2035

SAM · buyers we can serve now,
bottom-up

SOM · ARR by year 5: 3.2% of SAM

Who signs	Count	ACV	SAM	Yr-5 target	SOM
ML platform leads: own-model orgs	1,400+	\$75K	\$105M	30 teams (2.1%)	\$2.25M
Model-risk heads: 66 US + 110 ECB banks	176	\$250K	\$44M	10 banks (5.7%)	\$2.50M

Adjacent: AI governance software \$15.8B by 2030 (Forrester); MLOps \$16.6B by 2030. **All inside \$2.7T** of AI spend in 2026.
Funnel: next slide.

SNS Insider, AI MRM, 25 Aug 2026; Gartner, 16 Sep 2026; Fed large banks, 30 Jun 2026; ECB list, Jul 2026; W&B >1,400 orgs.
Forrester 2024; Grand View 2025; MarketsandMarkets 2025; Deloitte 2024 (2023 spend). ACVs and SOM: our hypotheses (A22-A25). All: A8.

Trackers chart training; Oxiedo records each fix

Our axis: a per-component record of each fix, made in training, on their machines.

Buyer's alternative	Type · threat	Layer state in training	Names the part	Resets it	Record of each fix	Validator workflow
Rewind to checkpoint and retry	Substitute · high	×	×	×	×	×
Trackers: W&B, MLflow	Indirect · med	✓	~	×	×	~
Research resets: ReDo, CBP (open)	Substitute · med	✓	✓	✓	×	×
Data quality: cleanlab (finds bad labels)	Indirect · low	×	×	×	×	×
MRM: ValidMind, ModelOp, Credo, Fiddler, Arthur	Adjacent · med	×	×	×	×	✓
Oxiedo (ORMAS)		✓	✓	✓	✓	~

The category is being bought: W&B → CoreWeave (~\$1.7B) · Arize → Dynatrace (\$915M) · Neptune → OpenAI. **And funded:** Arize \$70M, Fiddler \$30M, Credo AI \$21M, ValidMind \$8.1M (A10).

Our edge: a 383-experiment failure archive, data read in the loop, and a record that, once a validator accepts it, costs a re-validation to replace (A18–A19).

~ = partly; threat is our rating. Vendor docs, 13 Oct 2026. ReDo: ICML 2023. CBP: *Nature* 2024. Deals 2025–26: A10.

Ships fast, and tries to break the work first

Rokib Al Dhin Raadh, 18, Dhaka. Built V1 alone at 17.

DATED, CHECKABLE

Apr 2026	Own test suite killed the first design; rebuilt.
To date	383 controlled experiments, 4 model families.
Sep 2026	Replay demo shipped, oxiedo.com.
Sep 2026	DeepMath 2026 poster accepted.
Oct 2026	NeurIPS 2026 workshop reviewer.
Oct 2026	Ran every baseline reviewers asked for.
Earlier	AI-run e-commerce test: ~\$6.7K in 12 months, ~95% AI-referred, \$0 ads.*

SKILLS THE PLAN NEEDS

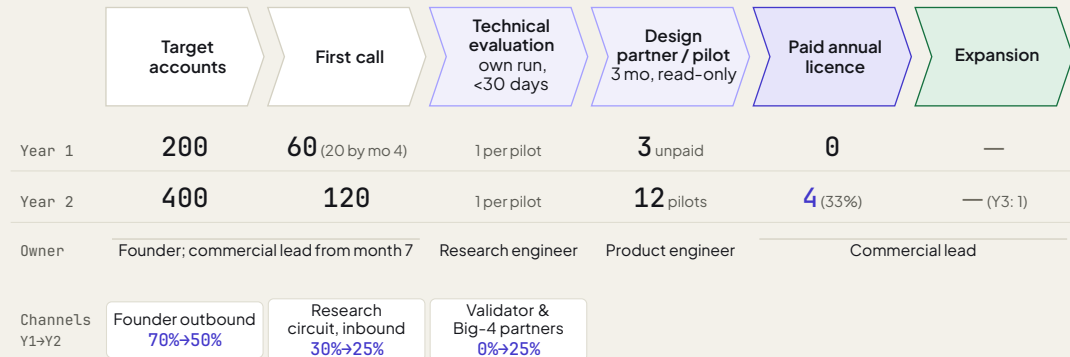
Skill		Evidence, or who covers it
Research	✓	ORMAS; 383 runs; DeepMath
Engineering	✓	V1 code; replay demo shipped
Product	~	Demo, site; + engineer M7
Sales	×	Commercial lead, month 7
Model risk	×	Bank MRM advisor, recruiting

Month 1: research engineer. Dhaka cost base; Delaware at close.

*Black Bloxie LTD (UK; father is director of record). One \$4,386 order is ~65% of revenue. Tested an earlier agent system, not ORMAS.

20 buyer calls by month 4; 4 paid teams in year 2

Every stage has a number, an owner and a channel. Each is a hypothesis until contract three.



All stage numbers are hypotheses, set before any contract. Month 4 = milestone M4. Funnel, unit economics, budget: A22–A24.

Five milestones, each with a pass line written first

Research runs only where the product needs an answer. A failed milestone gets reported.

Milestone	Passes if	Month
M1 Repair benchmark	5 seeds vs. 3 published fixes and a dropout control; ours within 1 point of the best	2
M2 Detection study	Injected faults found at $\geq 90\%$ precision and recall, flagged within 1 epoch	4
M3 Beyond CIFAR	M2's pass line holds on credit-risk and medical-image data	7
M4 Design partners	20 buyer conversations; 3 teams run the read-only monitor in-house	4 → 12
M5 Open record	Preprint and code public; NeurIPS 2027 submitted	5

Months from close, assumed Dec 2026; M5 is pinned to the NeurIPS 2027 deadline (mid-May). Tests: R4 §7; the reviewer asks they answer: A3.

\$750K buys five tested milestones in 18 months

Dhaka cost base. Delaware C-corp at close; relocation follows the first design partner.

\$750K

post-money SAFE

\$7.5M

cap · 10.0% dilution

18 mo

Dhaka cost base

Use of funds	%	Milestones
Research engineer, remote, month 1	22	M1-M3, M5
Product and commercial hires, month 7	30	M4
Compute (rented GPUs)	10	M1-M4
Founder, living in Dhaka	6	all
Delaware C-corp, IP assignment, legal	9	close, M4
Partner travel, conferences	5	M4, M5
Tools and insurance	3	—
Buffer, unallocated	15	—
Total \$750K	100	

Carta median cap for \$250K–\$1M SAFEs: ~\$10M (2025). Programme terms (EF/Antler) replace the instrument. Detail: A5, A11.

The system of record for what training did

Every model should arrive with a record of what its training did, part by part.

Rokib Al Dhin Raadh, Founder

raadxbusiness9@gmail.com

oxiedo.com · raadh.me

Replay the run: oxiedo.com/black-box

Appendix: the questions that come next

Grouped by the question an investor asks. Every entry is a link.

DOES IT WORK?

- [A1](#) What is not established
- [A2](#) Full results, every run
- [A3](#) Kill rules and venue record
- [A7](#) Every number and its source
- [A17](#) Scope: today, next, not at all

WHO BUYS, AND WHY?

- [A12](#) Personas and motivations
- [A13](#) Pains: obstacles to risks
- [A14](#) Gains at four levels
- [A15](#) Feature to value map
- [A16](#) Desired and unexpected gains

HOW BIG, AND AGAINST WHOM?

- [A8](#) Market sizes by analyst
- [A9](#) The buyer's pain, documented
- [A10](#) Category deals and pricing
- [A18](#) Competitor map by threat
- [A19](#) Why this is hard to copy
- [A20](#) Porter's five forces
- [A21](#) Price comparison

HOW DOES IT MAKE MONEY?

- [A4](#) Licence and pricing
- [A25](#) How we get paid
- [A22](#) Three-year acquisition plan
- [A23](#) Unit economics, CAC, LTV
- [A24](#) Go-to-market budget
- [A5](#) Three-year model, use of funds

THE DEAL

- [A6](#) Entity and IP
- [A11](#) Round benchmarks

A1: What is not established, and where ORMAS loses

NOT ESTABLISHED

Scope. Every result is CIFAR-10/100 on small CNN, MLP and ResNet models. No transformer, tabular or customer data.

Repair. One seed: plain reset 83.3%, our controller 78.5%, ReDo 77.5%. What is sold is the detector and the record; M1 re-tests on 5 seeds.

Detection. Precision, recall and delay against known faults are unmeasured (M2).

V2 signal. The CNN disagreement pilot was negative: AUROC ≈ 0.51 vs plain loss ≈ 0.97 .

Rebuild. The April suite that killed the predecessor has not been re-run on the current code.

Paper vs code. The paper states a 30% cap on resets; the code does not enforce it. Correction pending.

WHERE IT LOSES OR TIES

3 seeds, %	ORMAS	Plain
Random-direction injection	83.1	84.1
x100 weight explosion	85.1	86.0
ResNet-18 lesion	91.6	92.3
Clean CNN	85.6	85.1

Label noise 40%: ORMAS 75.6, CNN + dropout 74.5, plain 69.6. Most of the gain is dropout.

Cost: $\sim 2.6\times$ wall-clock (799 s vs 312 s); +15,867 parameters (+2.56%).

Register IDs: losses H-5; clean H-7; noise H-6; cost H-11/12; 1-seed baselines X-003; V2 pilot D-110/112; suite CP 21 risk 12; cap CP 21 V-1.

A2: ORMAS recovers what standard training cannot

Every archived lesion family, weak runs included.

Part zeroed at epoch 100	ORMAS	Comparison	Seeds	Reading
2 of 3 blocks, CIFAR-10 (headline)	80.3 ± 1.5 (79.9 / 81.9 / 79.0)	Standard 10.0, every seed	3	Repair . Standard is frozen: every gradient is 0
Same lesion, Oct 4 rerun	78.5	Plain reset 83.3 ; ReDo 77.5; standard 10.0	1	Plain reset beats ORMAS by 4.8 pp (dropout confound)
3 of 3 blocks, CIFAR-10	71.5 ± 2.5 (72.9 / 73.1 / 68.6)	Standard 10.0, every seed	3	Repair , partial
1 block + 40% label noise	31.8 / 76.9 / 69.4	Standard 10.0, every seed	3	~ Seed-variable; one seed reset the wrong block first
2 of 3 blocks, CIFAR-100	7.4 / 25.8 / 27.1 (pre-lesion 53.3)	Standard 1.0 (chance)	3	~ Weak; still rising at epoch 199

Reading. Every win is a part whose learning signal is exactly zero; recovery starts at the reset (9 of 9 runs, epochs 103–104). ORMAS carries Dropout(0.3), the plain reset none: M1 reruns row 2 with ≥ 3 seeds and matched dropout.

Rows 1, 3–5: archived runs (H-1, H-5; F3 D-070, D-071, D-075). Row 2: RTX 3080, PyTorch 2.8, seed 0 (FO X-003; D-131–134). 9 of 9: H-3. Wrong block first: D-065. All values: final test accuracy (%) at epoch 199.

A2: Clean, noisy and perturbed training

Clean data, label noise with the dropout control, perturbations and cost.

CIFAR-10 setting	ORMAS	Comparison	Seeds	Reading
Clean training	85.6 ± 0.4	Standard CNN 85.1 ± 0.5	3	~ Parity; epochs unequal
40% label noise, 200 epochs	75.6 ± 0.8	CNN + Dropout(0.3) 74.5 ± 0.4 ; plain 69.6 ± 0.9	5/3/3	~ Dropout explains ~5.0 of 6.0 pp; rest +1.05, $p \approx 0.05$
40% label noise, 100 epochs	—	CNN + Dropout(0.5) 76.6 ± 0.8	3	Stronger dropout alone is as good or better
Random-direction weight injection	83.1 ± 0.3	Standard 84.1 ± 0.3	3	Loss, -0.9 pp; below standard every epoch
x100 weight explosion	85.1 ± 0.5	Standard 86.0 ± 0.1	3	Loss, -0.9 pp; batch norm absorbs the shock
ResNet-18, blocks behind a skip	91.6 ± 0.4	Standard 92.3 ± 0.3	3	Loss, -0.7 pp; the skip keeps signal flowing
Cost of the lesion run	799 s	Standard 312 s; reset 342 s; ReDo 452 s	1	~2.6x wall-clock; +15,867 params (+2.56%)

Clean: H-7. Noise: H-6; F3 D-080/081 (Dropout 0.5: supp212). Injection (5x weight norm), explosion, ResNet: H-5, D-075.
Cost: RTX 3080, 1 seed (H-11, R-7); params on the 636,677-param CNN (H-12). Values: final test accuracy (%).

A3: We write the kill rule first, and follow it

Three pre-set tests went against us. Each one changed what we build.

When	Rule, written before the run	Outcome	What we did
Apr 2026	Predecessor architecture vs a purpose-built suite with a pass/fail list	Failed. Lost to a plain CNN at all four noise levels; zero corrections fired in 42 runs	Rewrote the system. The suite has not yet been re-run on the rebuild; that rerun is planned
Oct 3, 2026	V2 premise: disagreement between parts flags mislabelled samples (CNN pilot, 3 seeds)	Negative. AUROC ≈ 0.51 , vs 0.97 for plain per-sample loss	V2 moves to parallel parts. The next pilot must beat loss and AUM by ≥ 0.03 AUROC, published either way
Oct 4, 2026	Paper rule: if ORMAS ends 2+ pp below a plain reset, do not submit	Triggered. ORMAS 78.5 vs reset 83.3 (-4.8 pp, 1 seed)	Paper held. V2 keeps only detect + reset; the product is the detector and the record

Why this is here. A single-author result is easy to tune until it flatters. Each rule above was written first, went against us, and stays on the record.

Apr 2026 suite: Commercial Plan 21 §2B (killtest_results, 2026-04-10). V2 pilot: F3 D-110/112. Paper rule: FO X-003. Next-pilot bar: F4 S-149. AUROC: how well a score separates mislabelled from clean samples (0.5 = chance).

A3: Every reviewer ask maps to a milestone

Venue record: accepted at DeepMath 2026; reviewer for NeurIPS 2026.

Venue	Outcome	What reviewers asked for	Status milestone
DeepMath 2026	Accepted , poster (Sep 30); poster presented Oct 30. Short abstract applying known stability results to training-time repair	Simple re-init and ReDo baselines	Done , 1 seed (A2); M1 adds seeds + dropout control
NeurIPS 2026	Reviewer (Oct 3); Trustworthy AI for Good workshop, Paris (Dec 12)	CBP and checkpoint-rollback baselines	M1
		Leave-one-out ablations	Null on clean/noise (archived); under the lesion: M1
		Compute cost	Done : ~2.6x wall-clock (A2)
		Detection precision, recall, delay	M2
		Scale beyond CIFAR	M3
		Stability assumptions, framing, jargon, dropout claim	M5 preprint, rewritten around A2

Venues: OpenReview decision released Sep 30, 2026 (DeepMath #39); NeurIPS 2026 reviewer recognition letter, Oct 3, 2026.
Reviewer asks: F2 Part F (R-1 to R-16), F4 S-156. Leave-one-out nulls: F2 H-8. Milestones: plan slide.

A4: An annual licence per team, on its own systems

THE LICENCE

Annual, per team, run on the customer's own machines or cloud.

No seats, no usage meter, no hosted tier: training data never leaves them.

Wedge: monitor and change record for an ML platform team.

Expansion: the same record, read by a model-risk validator.

PRICE BAND: HYPOTHESIS

Monitor: \$75K/yr per ML team (band \$75–250K).

Model-risk record: \$250K/yr per validation team (band \$200–500K).

Set against what a failed run already costs the team and what model-risk tooling costs the validator.

No price has been put to a buyer yet; M4 tests it.

WHY IT RENEWS

Records belong to the customer. They export every record, in a documented format, during the term and after it.

Renewal is earned on **new runs**: monitoring, updates, support, and a version their validator has reviewed.

Switching costs a re-validation, not lost history.

Gross margin logic. Nothing is hosted, so serving cost sits with the customer. Cost of revenue is deployment engineering and support: modelled at 70% in year 2 and 80% in year 3, to be replaced by pilot data.

Bands: Commercial Plan 14 (monitor) and 16 (model risk); same as s08. Plan figures, untested. Margins are A5 assumptions.

A5: Three-year model, year 1 at \$0 revenue

\$, year-end	Year 1	Year 2	Year 3	Assumption
Design partners (unpaid)	3	3	—	M4: 3 teams on the read-only monitor by month 12
Paying monitor teams	0	4	9	\$75K/yr each; 1st converts by month 15
Paying model-risk teams	0	0	3	\$250K/yr each; validator expansion from Y3
ARR	0	0.30M	1.43M	Teams × licence (A4 bands, as s08)
Revenue in year	0	0.15M	0.86M	Contracts land mid-year on average
Gross margin	—	70%	80%	No hosting; cost is deployment and support
Headcount	4	9	16	Founder + 3 hires from this round (A5 detail)
Operating spend	0.40M	1.40M	2.60M	Y1 = this round's plan; then ~\$150K per loaded hire
Net burn	0.40M	1.30M	1.91M	Operating spend less gross profit
Money in	0.75M	3.0M	—	This round; seed at month 15 on M1–M5 (assumed)
Cash, year-end	0.35M	2.06M	0.14M	Series A in year 3; not modelled

Unit economics (CAC, LTV, payback): A23, as hypotheses until the first three paid contracts replace them.

Every figure is a planning assumption, not a forecast from data. Year 1 = months 1–12 after close. Licence band: A4.

A5: The \$750K, line by line

Line	Mo 1–12	Mo 13–18	Total	%	Assumption → milestone
Research engineer, remote, month 1	110.0	55.0	165.0	22	\$9.2K/mo loaded → M1–M3, M5
Commercial lead, month 7	60.0	60.0	120.0	16	\$10K/mo → M4
Product engineer, remote, month 7	52.5	52.5	105.0	14	\$8.75K/mo → M4
Compute, rented GPUs	50.0	25.0	75.0	10	Non-CIFAR data, pilots → M1–M4
Founder, living in Dhaka	30.0	15.0	45.0	6	\$2.5K/mo → all
Delaware C-corp, IP, legal, accounts	35.0	10.0	45.0	6	Formed at close → M4 contracts
Regulatory, data-protection review	22.5	0.0	22.5	3	One engagement each → M4
Partner travel, conferences	25.0	12.5	37.5	5	Partner visits, NeurIPS → M4, M5
Tools, software, insurance	15.0	7.5	22.5	3	—
Buffer, unallocated	—	—	112.5	15	Slippage, or a 3-month bridge
Total, \$K	400.0	237.5	750.0	100	

Seed timing. The seed is raised from month 12 on M1–M5 results (A5 model assumes month 15). The buffer covers three more months if it slips.

Current experiments cost about \$200–600 of rented GPU time; the compute line budgets for non-CIFAR data, larger models and partner pilots.

A6: Oxiedo incorporates in Delaware at close

TODAY

Oxiedo is not yet incorporated.

No cap table, no prior investor.

The founder personally holds the ORMAS code, run archive, paper and brand.

Black Bloxie LTD is a separate UK company; the founder's father is director of record. It ran the e-commerce experiment and is the interim contracting party for oxiedo.com.

AT CLOSE

Delaware C-corp formed before the SAFE is executed; cost is in the legal line (A5).

IP assignment at incorporation: code, archive, paper, brand and domain.

Any Oxiedo agreement or right held by Black Bloxie LTD moves to the C-corp; Black Bloxie takes no equity.

Founder stock vests. The founder, 18, is the contracting party.

CODE AND LICENCE

The code is not public today.

The architecture and run archive are **released on publication** under a research licence: free for research, teaching and evaluation. Production use is licensed (A4).

Not open source. An earlier MIT notice in the private repository is replaced before release.

Entity: README; oxiedo.com/about, /press. Licence: oxiedo.com/licensing. Bangladesh FX rules on US shares: counsel review before close.

A7: Every number and its source

Measured figures come from the run archive; market and price figures are labelled hypotheses.

Figure (slide)	Source	Status
10.0% vs 80.3 $\pm$ 1.5% (3, 4) Stuck at chance 12/12; reset at epoch 103–104 in 9/9 (3, A2) Plain reset 83.3, ReDo 77.5, ORMAS 78.5 (4, A2) 383 experiments (10) ~20 loss spikes (2) EU 2 Dec 2027; SR 26–2; SS1/23 (7) ~2.4x/yr training cost (7) \$5.9B (2025) $\rightarrow$ \$19.1B (2035) (8) 1,400+ orgs; 176 banks; \$149M (8) \$170M; 35+ restarts (2) 365 models; 29% (7) \$1.7B; \$915M deals (9) Competitor capabilities (9) ~\$6.7K, ~95% AI-referred (10) \$750K, \$7.5M cap, use of funds (12, A5)	Archived lesion runs, 3 seeds each; F2 H-1, F3 D-070/071 F2 H-2, H-3; archive-wide recount Rerun, RTX 3080, PyTorch 2.8, seed 0; FO X-003 Run records across the founder's result folders (617 run IDs on disk) PaLM, arXiv:2204.02311 \$5.1 Reg. (EU) 2026/1744; Fed SR 26–2; PRA SS1/23 Cottier et al. (Epoch AI), arXiv:2405.21015 SNS Insider, AI model-risk management, 25 Aug 2026 W&B customer base; Fed large banks; ECB list; ACVs ours Stanford AI Index 2025; OPT-175B logbook (Meta, 2022) Deloitte EMEA model-risk survey 2025 CoreWeave, Dynatrace announcements Vendor documentation, read 13 Oct 2026 Black Bloxie LTD store records, 12 months Carta pre-seed data 2025–26; our plan	Measured; rerun 4 Oct 2026 Measured Measured, 1 seed; M1 repeats it Counted Third-party paper Primary regulatory text Third-party paper Analyst report Counts sourced; prices untested Third-party Third-party survey Public filings/press Public claims Founder records Proposal

Registers: oxido/FINDINGS (FO conversation, F2 paper claims, F3 data). Full results table: A2. Limits: A1.

A8: Analysts size the market in billions

AI model-risk management, its adjacent markets, and the spend around them.

Market	Size	Source, date
AI model-risk management	\$5.87B (2025) → \$19.10B (2035) , 12.5%/yr; finance 35%	SNS Insider, 25 Aug 2026
AI model-risk management	\$5.7B (2024) → \$10.5B (2029), 12.9%/yr	MarketsandMarkets, Jul 2024
AI model-risk management	\$6.6B (2025) → \$16.0B (2032), 13.5%/yr	P&S Intelligence, Aug 2026
AI model-risk management	\$7.17B (2025), \$8.33B (2026), \$15B (2030)	The Business Research Co., Sep 2026
AI governance software	\$15.8B by 2030 , ~4× 2024	Forrester, 2024–25
AI governance platforms	\$492M (2026) → >\$1B (2030)	Gartner, 17 Feb 2026
AI trust, risk & security mgmt	\$3.09B (2026) → \$11.61B (2031), 30.3%/yr	MarketsandMarkets, Aug 2026
MLOps	\$2.19B (2024) → \$16.61B (2030), 40.5%/yr	Grand View Research, 2025
Data collection & labelling	\$17.1B by 2030, 28.4%/yr	Grand View Research, Nov 2024
AI in healthcare	\$21.66B (2025) → \$110.61B (2030)	MarketsandMarkets, 2025
AI in military	\$14.3B (2024) → \$29.0B (2030)	Research and Markets, Jun 2025
Worldwide AI spend	\$2.7T (2026) , +49.5%; infrastructure \$1.43T	Gartner, 16 Sep / 19 May 2026
Big-4 hyperscaler capex	~\$700–725B (2026), from ~\$410B	Company guidance, Q1 2026
Top-20 pharma R&D	\$145.5B (2023 spend)	Deloitte, 2024

Vendor market reports differ in scope; the deck's headline uses one series (SNS Insider) end to end. URLs: [rework/research/R6](#).

A9: The buyer's pain is documented

Banks carry hundreds of models; large training runs fail as a matter of course.

Fact	Figure	Source, date
Models per bank (EMEA)	small 58 · medium 132 · large 365	Deloitte MRM survey, Oct 2025
Banks using AI/ML models	67% (90% of large); only 29% say their framework fits	Deloitte, Oct 2025
AI validation staffing	75% of banks put <5% of staff on AI validation	Deloitte, Oct 2025
Models per bank (US)	mean 175; cost (69%) and talent (56%) block more validation	RMA MRM survey, 2024
Cost to validate one framework	\$75,000 or more (external)	RMA, Nov 2021
Validation time	1–17 weeks (US); up to 30 weeks (Europe)	McKinsey, ~2017
Frontier training cost	Llama 3.1–405B ~\$170M; GPT-4 ~\$79M; Gemini Ultra ~\$192M	Stanford AI Index 2025
Cost growth	2.4× a year since 2016; \$1B runs by 2027 on trend	Epoch AI, 2024
OPT-175B	35+ manual restarts, 100+ hosts cycled in 2 months	Meta logbook, 2022
Llama 3 405B	419 unexpected interruptions in a 54–day window (>90% effective)	Meta, arXiv:2407.21783
ByteDance MegaScale	job repaired and recovered >100 times; failures “the norm”	NSDI 2024
EU AI Act fines	€35M/7% prohibited; €15M/3% high-risk	Reg. 2024/1689 Art. 99
Risk-governance penalties	Citi: \$400M (2020), \$136M (2024), data and risk governance	OCC, Fed

Citi penalties concern risk and data governance, not model risk specifically. URLs: [rework/research/R7](#).

A10: Buyers are paying for this category now

Training and model observability is being acquired and funded, and priced in five figures.

Deal or price	Value	Source, date
CoreWeave acquires Weights & Biases	~ \$1.7B (reported); closed May 2025	CoreWeave; TechCrunch
Dynatrace to acquire Arize	\$915M	Dynatrace, 13 Aug 2026
OpenAI acquires Neptune (training-run tracking)	undisclosed; reported <\$400M in stock	OpenAI, Dec 2025
Cisco acquires Robust Intelligence; Galileo	undisclosed	Cisco, Sep 2024; Apr 2026
Snowflake acquires TruEra	undisclosed	Snowflake, May 2024
Arize Series C	\$70M	TechCrunch, Feb 2025
Fiddler Series C	\$30M (~\$100M total), regulated industries	Jan 2026
Credo AI	\$21M (\$41.3M total)	SiliconANGLE, Jul 2024
ValidMind seed, bank model risk	\$8.1M, led by Point72 Ventures	FinTech Global, Mar 2024
W&B contract, median	\$42,900/yr (52 purchases; \$6.3K–\$136K)	Vendr, 2026
ValidMind	~\$5,000 per model per year	AWS listing (secondary)
Credo AI	annual, per AI use case	AWS Marketplace
External validation, one framework	\$75,000 or more	RMA, 2021

Our planned ACVs (\$75K monitor, \$250K model risk) sit between tool pricing and validation spend: A4. URLs: rework/research/R7.

A11: The round is priced inside market norms

\$750K at a \$7.5M cap sits below Carta's median for its size.

Benchmark	Value	Source, date
Median cap, \$250K–\$1M SAFEs	~\$10M	Carta, State of Pre-Seed 2025
Median cap, \$1M–\$2.5M SAFEs	~\$15M	Carta, State of Pre-Seed 2025
Median cap, SAFEs >\$2.5M	\$35M, +40% year on year	Carta, Q2 2026
Pre-seed rounds on SAFEs	93%	Carta, Q2 2026
Post-money SAFEs with a cap	94%	Carta, H1 2026
Cap-only SAFEs (no discount)	61% (8,762 SAFEs, 2024)	Carta
Median US pre-seed deal	~\$0.9M	PitchBook–NVCA via SaaStr, Q2 2025
SAFE volume, 2025	\$10.4B across 50,316 instruments	Carta, Feb 2026
AI valuation premium	1.2–1.5× pre-money, seed to Series C	Gunderson Dettmer, 2025
Solo founders	36.3% of new startups (H1 2025)	Carta Solo Founders Report
YC	\$500K: \$125K for 7% + \$375K uncapped MFN	Y Combinator
Antler Singapore	\$150K: \$100K for 10% + \$50K MFN	Antler, 2026
Entrepreneur First	\$250K: \$125K for 8% + \$125K MFN	third-party, 2026

Programme terms replace s12's instrument; milestones are unchanged. URLs: [rework/research/R8](#).

A12: Three people decide, each for a reason

	P1 ML platform lead	P2 Head of model risk	P3 Model validator
Company	Trains its own models: AI-native, fintech or bank data-science team (1,400+ orgs, W&B base)	Bank over \$30B in assets (66 US) or ECB-supervised (110)	Bank's second line, or a Big-4 validation practice
Responsibilities	Training infrastructure, compute budget, run reliability	Model inventory (large bank $\approx$ 365 models), validation programme, exams (SR 26-2, SS1/23, EU AI Act)	Independently reproduce evidence; write findings
Current workflow	W&B / MLflow dashboards; on failure, rewind to a checkpoint, skip data, retry	Collect developer documents; 1-17 weeks per validation; \$75K+ per external framework validation	Read docs, rerun tests, request evidence, which is often missing for training
Buying authority	Signs tool budgets around \$75K/yr (<i>hypothesis</i>)	Owens the MRM budget; \$250K needs CRO sign-off (<i>hypothesis</i>)	No budget; can block (gatekeeper)
Functional	Fewer lost runs; know which part failed	Pass validation first time; cover AI models	Verifiable, reproducible evidence
Emotional	Stop firefighting at 3am	Not be the name on a reopened model	Confidence in a sign-off
Social	Platform team seen as reliable	Clean exam record for the bank	Defensible findings with the regulator

Counts: W&B base, Fed (30 Jun 2026), ECB; 365 models: Deloitte 2025; weeks: McKinsey; \$75K+: RMA 2021. Authority: hypothesis.

A13: Each persona's pain, from obstacle to risk

The obstacle is the same for all three: nobody can say what training did to which part.

	P1 ML platform lead	P2 Head of model risk	P3 Model validator
Obstacles	Dashboards show the loss, not which component failed. Fixes are trial and error: OPT-175B needed 35+ manual restarts in 2 months; PaLM's loss spiked ~20 times	A large bank runs 365 models ; 75% of banks put <5% of staff on AI validation. Training evidence arrives as prose, if at all	Receives a finished model. Cannot replay what training did, or which interventions happened and why
Undesired outcomes	Rewind to a checkpoint and skip batches with no named cause; GPU-hours and deadlines lost; the same failure recurs	Validation takes 1–17 weeks per model and \$75K+ per external framework validation; AI models queue	Repeated evidence requests to developers; findings delayed or written on incomplete evidence
Risks	A silent failure ships a weaker model; at frontier scale a run costs \$79–192M	Exam findings on AI models; EU AI Act fines up to €15M or 3% for high-risk breaches; a reopened model	Sign-off a regulator later reverses; findings that cannot be defended

Meta OPT logbook 2022; PaLM arXiv:2204.02311; Deloitte 2025; McKinsey; RMA 2021; AI Index 2025; Reg. 2024/1689 Art. 99.

A14: Gains at four levels, with how we measure them

Each gain has an indicator; **T** = target set by a milestone, **H** = hypothesis tested in pilots.

	P1 ML platform lead	P2 Head of model risk	P3 Model validator
Required minimum to buy	Dead component named in the record within 1 epoch of failing → $\geq 90\%$ precision and recall on injected faults (M2) T	Record accepted as validation evidence → validator accepts it in one validation cycle (contract trigger) T	Reproducible evidence of what training did → a replay returns the same component, epoch and change T
Expected	Run recovers without a rewind → share of flagged runs recovered without restarting from a checkpoint H	Shorter validation cycle → weeks per AI-model validation, against the 1-17-week baseline H	Fewer evidence requests to developers → requests per model validation H
Desired	Fewer lost GPU-hours → GPU-hours lost to failed runs per quarter, vs the team's prior quarter H	Clean exam findings on AI models → open supervisory findings on in-scope models H	Faster, defensible sign-off → days from evidence received to findings issued H
Unexpected upside	Reusable failure history across runs → past failures searchable by component and cause H	Lower re-validation cost on retrains → cost per re-validation, against \$75K+ external H	Failure patterns to test new models against → findings that cite a known pattern H

Measured base today: dead blocks flagged at epoch 101 after a lesion at epoch 100; recovery starts at the reset in 9/9 runs (F2 H-3).

A15: Every feature maps to a pain and a value

Three features exist in research code today; packaging and export are milestone work.

Feature	Status	Pain resolved (persona)	Customer outcome	How quantified
Detection	Today, CIFAR scale	P1: nobody can say which part failed	The dead component is named while the run trains	Detection precision, recall and delay on injected faults (M2)
Targeted reset	Today; opt-in; prior art (ReDo, CBP)	P1: rewind, skip data, retry	The run recovers without a rewind	Accuracy recovered: 80.3% vs 10.0% stuck (lesion test)
Change ledger	Today	P2, P3: training evidence missing or prose	What changed, where, when and why, on file	Fields per intervention: component, epoch, reason, size
Read-only deployment	Planned, M4	P1: will not let a tool change training; P2: data stays in-house	Runs on their machines; observes and records only	3 teams run it in-house (M4); data leaving their systems: none
Validator export	Planned, M4	P2: 1-17 weeks per validation; P3: cannot reproduce	A record a validator can read and replay	Record accepted in one validation cycle; weeks saved (<i>hypothesis</i>)

80.3 $\pm$ 1.5% vs 10.0%: 3 seeds, CIFAR-10 lesion (F2 H-1). The reset itself is prior art; the product is detection plus record.

A16: Gains they ask for, and gains they get anyway

Buyers ask for recovery and audit readiness; the record also builds a history.

DESIRED: WHAT THEY ASK FOR

Faster recovery. The failed component is named and reset; the run continues instead of rewinding.

Audit readiness. Every intervention is on file, by component and epoch, before the validator asks.

Fewer lost runs. Failures are caught while training, not found at evaluation.

Measured by: detection delay (M2), weeks per validation, GPU-hours lost per quarter.

UNEXPECTED: WHAT THEY GET ANYWAY

Reusable training history. Past failures stay searchable by component and cause across runs.

Lower re-validation cost on retrains. The validator reviews what changed, not the whole model again.

A failure library across teams. One team's failure becomes another team's known pattern.

Measured by: re-validation cost vs \$75K+ external; reuse of known patterns (*hypotheses*).

Unexpected gains are hypotheses for pilots (M4). \$75K+ per external framework validation: RMA 2021.

A17: What ORMAS covers today, next, and not at all

Today is dead components on small vision models; each extension is tied to a milestone.

	Today	Planned	Not supported
Failure types	Dead layers and units: detect, reset, record (CIFAR)	Loss spikes and divergence; injected-fault detection with precision/recall (M2); label noise (V2)	Inference-time drift; deliberate attacks on a model
Model families	MLP, CNN, ResNet (CIFAR scale)	Tabular/credit and medical images (M3); transformers (post-seed)	LLM pre-training at frontier scale
Workflows	PyTorch training loop; replayable record	Packaged read-only monitor; validator export (M4)	Hosted service (by design); MRM-platform integration (later)

Where nothing is broken, ORMAS ties or trails, at $\sim 2.6\times$ wall-clock cost (A2). That is why the product is the monitor and the record, not a training replacement.

Cost: 799 s vs 312 s wall-clock, +2.56% parameters (F2 H-11, H-12); clean and loss results: A2. Milestones M2-M4: s11.

A18: One map of every alternative, typed by threat

Substitutes are the real rival: rewind and retry. None of these writes a per-component fix record.

Alternative	Type · threat	Strength	Weakness vs Oxiedo	Oxiedo advantage
Rewind and retrain	Substitute · high	Free; in every team's runbook	Repeats compute; no cause	Names the part; records it
W&B (CoreWeave)	Indirect · med	Layer histograms; 1,400+ orgs	Charts it; no fix, no record	Acts in the loop; records the fix
MLflow (open source)	Indirect · med	Free; metrics, model registry	Logs only what you send	Adds the diagnosis and record
Neptune (OpenAI)	Indirect · low	Per-layer metrics at GPT scale	External service closed 2026	Its former users need a tool
Arize (Dynatrace)	Indirect · low	Production ML/LLM monitoring	Watches deployment only	Covers the training phase
Fiddler	Adjacent · med	Explainability, monitoring	Post-deployment only	Evidence from inside training
Arthur	Indirect · low	Production monitoring, evals	Post-deployment only	Evidence from inside training
Robust Intelligence	Indirect · low	Model security; now Cisco	Tests the finished model	Records what training did
ValidMind	Adjacent · med	Bank MRM validation docs	Needs developer evidence	Supplies that training evidence
ModelOp, Credo AI	Adjacent · med	Inventory, policy, governance	Process, not model internals	A record they can reference
cleanlab	Indirect · low	Finds bad labels and data	Data, not network parts	Complementary; labels are V2
ReDo, CBP (open code)	Substitute · med	Peer-reviewed; same reset	No record, no product	Same reset, plus the record
In-house scripts	Substitute · med	Fits their stack; no vendor	No format a validator knows	A standard record format

Threat = our rating. Vendor docs, 13 Oct 2026; R7 rows 33–49. Neptune: external service wound down Mar 2026. Dynatrace–Arize: agreed Aug 2026.

A19: Why this is hard to copy

Five advantages, each with what a competitor would need, and what we have today.

Advantage	What it is	What a competitor needs	Status today
1 The failure archive	383 experiments; a labelled taxonomy: what dies, how it shows, what fixes it	The same controlled experiments, run and labelled from scratch	Built; grows with each design partner (with consent)
2 Inside the training loop	Reads component state while training runs	A hook in the loop; post-hoc tools cannot recover it later	Built for PyTorch; packaged monitor is M4
3 Validator-accepted format	Once a validator accepts the method, replacing it means re-validating	Their own validation cycle, bank by bank	Not yet: no validator has reviewed it
4 Regulatory evidence	Maps to EU AI Act Annex IV 2(d), SR 26–2 ongoing monitoring, FDA PCCP	Rule-by-rule mapping plus a record that fits it	Mapping stated; untested with a regulator
5 Speed and discipline	Pre-registered tests, and the record of following them (A3)	A track record; cannot be bought	Kill rules written before runs (A3)
Not a moat	The reset itself: prior art (ReDo, ICML 2023; CBP, <i>Nature</i> 2024)	Nothing: the code is open	A plain reset matches ORMAS (A2)

Honest ranking: 1 and 2 exist now. 3 is what locks in, and it is earned in a first validation cycle, not claimed. A tracker could copy 2 quickly; 1 and 3 take runs and a validator's review.

Moat list: rework/01_DEPTH_SPEC. Annex IV: Reg. (EU) 2024/1689. SR 26–2: Fed, 17 Apr 2026. PCCP: FDA final guidance, Dec 2024.

A20: Five forces: substitutes are the real rival

Rated by us. The plan answers the two high forces first.

Force	Rating	Why	Our response
Suppliers (GPU, cloud)	low-med	We host nothing; GPU prices hit the buyer's training bill, not our cost of sales	Runs on their machines; our research runs use \$2-4/h neocloud H100s
Buyer power	high	Few large buyers (176 banks, 1,400+ ML orgs); free options; the validator can block a sale	Free pilot on their own run; priced at one external validation (\$75K+)
Rivalry	medium	Funded adjacent players (Arize \$70M, Fiddler \$30M); none writes a per-component fix record (A18)	Stay on the training-time record; partner with trackers and MRM platforms
Substitutes	high	Rewind-and-retry is free and habitual; ReDo and CBP code is open	Sell the diagnosis and record, not the reset; show time saved in the pilot
New entrants	medium	W&B logs per-layer histograms; OpenAI took Neptune in-house; a reset is easy to add	Archive and validator acceptance take time; publish the method first (M5)

Round focus: pilots on the buyer's own run answer substitutes (M2, M4); validator acceptance answers buyer power and becomes the switching cost.

Ratings are our judgement. Counts: Fed Jun 2026, ECB Jul 2026, W&B base (R6). H100 \$/h: GetDeploying, Jun 2026 (R7 row 10).

A21: Priced between tooling and validation spend

Per customer per year. Our prices are hypotheses, untested with a buyer.

Alternative	List or observed price	Per customer per year	Source
W&B (tracker)	52 deals; range \$6.3K–\$136K	\$42.9K median per org	Vendr, 2026
Arize AX (production)	Pro \$50/mo; enterprise custom	\$0.6K self-serve; enterprise unpriced	Arize docs, 2026
ValidMind (MRM)	~\$5K per model per year	50 AI models ~ \$250K ; 175 models ~\$875K	Secondary (low confidence)
External validation	\$75K+ per framework validation	\$75K+ for one framework a year	RMA, Nov 2021
Rewind and retrain	Engineer time + repeat compute	Unpriced per team; frontier run \$79–192M	Stanford AI Index 2025
Fiddler, Arthur, ModelOp	No public price found	–	R7 row 38
Oxiedo monitor	Annual licence per ML team	\$75K (band \$75–250K)	Hypothesis
Oxiedo model-risk record	Annual licence per bank	\$250K (band \$200–500K)	Hypothesis

The price gap: the monitor is ~1.75x the W&B median because it fixes and records, not just charts; it costs about one external validation. The bank record matches a 50–model ValidMind scope, well under a 175–model inventory.

Assumptions: 1 framework validated a year; 175 = RMA mean US inventory. ValidMind figure: dupple.com, unverified. Full table: R7 §5.

A22: Three-year acquisition plan

Hypothesis. Each year's counts produce A5's paid customers: 0, then 4, then 12.

Stage	Year 1	Year 2	Year 3	Note
Target accounts worked	200	400	800	P1 organisations and P2 banks (A12)
First conversations	60	120	250	First 20 by month 4 = M4
Design partners / pilots	3 unpaid	12 pilots	30 pilots	Pilot: 3 months, read-only
New paid (conversions)	0	4	8	Y2: 33% of pilots; Y3: 5 monitor + 3 model-risk
Total paid	0	4	12	= A5: 9 monitor + 3 model-risk in Y3
Expansion	—	—	1 team	A monitor customer adds a 2nd team (in the 9)
Mix: founder outbound	70%	50%	40%	Founder in Y1; commercial lead from month 7
research circuit, inbound	30%	25%	25%	Papers, NeurIPS, replay demo
validator, Big-4 partners	—	25%	35%	Route to P2 banks through P3
S&M spend, fully loaded	\$120K	\$220K	\$500K	Total \$840K; detail A23, A24
ARR, year-end	\$0	\$0.30M	\$1.43M	= A5

Hypotheses; no contract yet. Replaced by observed conversion rates after the first 3 contracts (A23).

A23: Unit economics, until contracts replace them

CAC by channel	CAC
Outbound (founder, lead)	\$90K
Research circuit, inbound	\$50K
Validator, Big-4 partner	\$60K

CAC by deal	
Monitor deal (P1)	\$55K
Bank model-risk deal (P2)	\$120K

Blended CAC, how computed	
S&M Y1-Y3: 120 + 220 + 500	\$840K
÷ 12 paid by end of Y3	≈\$70K

Hypothesis	Monitor	Model-risk	Blended, Y3
ACV	\$75K	\$250K	≈\$119K
Gross margin	80%	80%	80%
Customer life	4 yr	5 yr	4 yr
LTV	\$240K	\$1.0M	≈\$381K
CAC	\$55K	\$120K	\$70K
LTV / CAC	4.4x	8.3x	≈5.4x
Payback	~14 mo	~10 mo	~12 mo

Monitor at the blended \$70K CAC: 3.4x. Benchmark: 3x is healthy. Payback = CAC / monthly gross profit + 3 months' ramp (monitor: 11 + 3).

Update rule. After contracts 1-3, recompute CAC and payback from actual cost; drop any channel above \$120K CAC.

Hypotheses; no paid contract yet. CAC = fully loaded S&M / paid customers won. LTV = ACV × margin × life.

A24: Each go-to-market dollar has an expected result

Line (A5 use of funds)	Y1 \$K	18 mo \$K	Drives	Expected result
Founder time (30%)	9.0	13.5	Outbound, months 1–6	20 conversations by month 4 (M4); 42 outbound in Y1
Commercial lead, from month 7	60.0	120.0	Outbound, pipeline, close	200 accounts in Y1; first paid monitor by month 15
Product engineer (50%)	26.25	52.5	Design-partner pilots	3 design partners on the monitor by month 12 (M4)
Partner travel, conferences	25.0	37.5	Research circuit, partners	18 research-circuit conversations in Y1; partners in Y2
Research engineer (evaluations)	—	—	Evaluation in each pilot	In R&D, not S&M
Go-to-market total	120.25	223.5	= Y1 S&M (A23)	Y1: 60 conversations, 3 design partners, \$0 revenue

After year 1. Y2: \$220K S&M → 4 new paid (\$55K each), \$0.30M ARR. Y3: \$500K → 8 new paid (\$62.5K each), \$1.43M ARR. Blended CAC \$70K.

Hypotheses. \$223.5K of the \$750K is go-to-market. Y1: \$2K per conversation, \$40K per design partner. Time shares estimated.

A25: How we get paid

Two annual licences, paid upfront. A pilot becomes a contract on a test the buyer agreed to first.

	Training Health Monitor	Model Change Record
Buyer	P1, ML platform lead	P2, head of model risk (P3, the validator, must accept)
Contract trigger	The pilot finds an injected fault in their own run, inside the agreed delay	The validator accepts the record in one validation cycle
Annual licence	\$75K per team (band \$75–250K)	\$250K per bank (band \$200–500K)
Payment	Annual, upfront	Annual, upfront
Pilot → paid	Y1: design partner, 3 months free. From Y2: paid pilot \$15K, credited to year 1. 33% pilot→paid	Starts from a monitor customer, or a 6-month validation pilot at \$40K, credited
Implementation, support	Implementation \$10K one-off; support included	Implementation \$25K; assurance work priced per engagement

Hypotheses until the first 3 contracts. Prices as A4, s08. Personas P1–P3: A12.