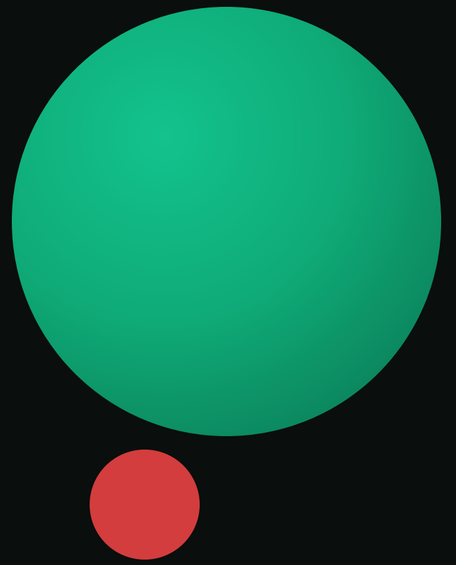


Go X OpsGym. The real world, made *trainable*.

An RL environment for physical-economy AI, seeded from a real fleet of 20,000+ scooters across three cities. Scooters is the first body. Every real-world business is a training environment.

Alex Debelov — product · Vlad Chistiakov — architecture · Khodor Salam — ops

SECTION 01 · THE GAP

Labs are already paying for reality.

Just the *dead* kind.

Every business-ops benchmark today runs on mock or historical data. Go X runs 20,000+ scooters across three cities with multi-city permits, impounds, field crews, tickets and a real P&L. That is the gap.

ANTHROPIC · 2026

$1B+

Reported single-year discussion on RL-environment spend. Environments are a budgeted line, not a science project.

MERCOR · JULY 2026

$2B

Gross run rate, doubled in four months. RL environments are its fastest-growing line, sourced mostly from five labs.

SPIRIT AIRLINES · AUG 2026

$10M

Google's winning bid for a bankrupt airline's dead archive — 100M emails, 500M Teams messages, code, ops records.

PROTEGE · 2026

$100M+

2026 data transactions, up from about $30M the prior year. Data owners license, labs consume.

SIMPLECLOSURE · 2026

~100 deals

Dead startups' Slack, email and code sold at $10K–$100K each. Reality has a market. Live reality has no supply.

THE GAP

Live judgment.

Agent acts, money grades. No price yet. No supply. That is the row we build.

The Information via Epoch AI (Jan 2026); Bloomberg Law, Axios, Forbes, TIME (Aug 2026); Sacra and The Information on Mercor; Forbes (Apr 2026); Epoch AI RL environment FAQ (Sept 2025). See RL Gym Playbook slides 4–5.

SECTION 02 · THE WEDGE

Nobody sells *live, outcome-verified* decisions from a physical operating business.

Computer gyms clone Salesforce. Robot-data startups train pick-and-place. We sit one level up: AI to economic objective to thousands of software, human, and machine actions — graded by the economy.

COMPUTER GYMS

Deeptune, HUD, Mercor, Scale

Cloned Salesforce and spreadsheets, rubric graders, resettable containers. Mercor bought Deeptune in July 2026. The computer version is already priced.

ROBOT DATA

Praxis, Enact, Claru

Robot to physical task. Pick up the box, fold the shirt, recreate the failure, train the recovery. Valuable — and one level too low for operating decisions.

MIDAS · OPSGYM

Go X and every business after it

AI to *economic objective* to thousands of software, human and machine actions. Run this fleet. Operate this cluster. Restock this store. The economy grades it.

Go X is the *first body* the brain runs. The pattern generalizes to every physical-economy business — cleaning, moving, storage, ghost kitchens, home services. `Fifty-two shapes and counting.`

Midas pitch, slides 1–3 (Sept 2026); Midas 52-shape catalog; TechCrunch and Forbes on Deeptune → Mercor acquisition (Jul 9, 2026); YC 2026 batch on Praxis/Enact.

SECTION 03 · THE 100-TASK LEDGER

One ledger. Three buckets. Three products.

Every row in the Go X 100-Task Ledger sorts into one of three lanes. Each lane is a different thing to sell.

BUCKET · A ROWS

10–15

Autonomous today

Rows the agent already runs without a human on the button. **This is the public number.** Ships as the OpsBench headline, updated monthly.

BUCKET · A+H ROWS

30–35

Agent proposes, human approves

Every run yields a labeled example: proposal, correction, verified outcome. **A+H rows are the trajectory archive** — a labeled example every time it runs.

BUCKET · H+JUDGMENT ROWS

50–60

Human decides today

Rows blocked on judgment where **no lab has a verifier**. Rows 7, 26, 43, 51, 72, 80 flagged as most valuable. **These become the environment** — labs pay to have it.

WHAT A+H BECOMES

A **trajectory archive**: real proposals, real corrections, real outcomes at t+1 / t+7 / t+30. This is the data-product line — sold, not open-sourced.

WHAT H+JUDGMENT BECOMES

A **resettable environment** with graders labs can run. The rows nobody else has a verifier for become the rows we ship one for.

Go X 100-Task Ledger, draft v0 (Sept 2026); GoX OpsGym Demo Spec §0 — "The decision this spec makes" (Sept 2, 2026).

SECTION 04 · WHAT WE BUILD

Three things, wired together, shown in one sitting.

Not a digital twin of Go X. An environment, two task tiers on top of it, and a provenance archive that grounds it.

01 · WORLD

Fleet Ops World

A discrete-event sim of one Go X city, exposed to an agent via HUD v6. MCP tools, workspace files, deterministic per `(scenario_id, seed)`. Identities synthetic, distributions real.

`gox-fleet-ops` `HUD v6` `Docker` `MCP`

02 · TASKS

Two tiers on top

Six short-horizon families drawn from the Judgment rows (5–15 tool calls each), plus one long-horizon "run a venue cluster for 30 days" in the Vending-Bench 2 shape that becomes the OpsBench headline.

`F1–F6` `30-day` `OpsBench` `group=16`

03 · PROOF

Provenance archive

A small, de-identified archive of real A+H decisions from Go X ops — each linked to the sim task minted from the same situation. 30–60 records for the demo.

`A+H records` `sim_scenario_id` `NDA sample`

Together this is *the world, its tests, and the receipts* — the three things a lab reviewer needs to see in one meeting to say "yes, we want to run this."

SECTION 05 · FLEET OPS WORLD

One city. Everything the ops team sees.

Distributions are real Go X. Identities are synthetic. That combination is what makes the world both defensible and clean.

CITY 1	VENUES 24	SCOOTERS 340	TECHS 4	BATTERIES 460	FAULT CODES 37	IMPOUND LOTS 3	TICKETS/DAY ~40

LEDGER per-city P&L	INBOX GMs, suppliers, clerk, riders	DEMAND rides × venue × hour × weather	CONTRACTS tiered commissions

TIME · SNAPSHOT MODE

Short-horizon tasks. Sim frozen at a moment (06:00 Friday before an event weekend). Agent acts; grader forward-simulates `N` days under a fixed baseline to measure the outcome.

TIME · CONTINUING MODE

The 30-day scenario. Agent calls `sim.advance_day()`. Overnight: inbox refreshes, demand realizes, batteries drain, faults and impounds occur, fees post. Vending-Bench pattern.

DETERMINISM

Same `(scenario_id, seed)` ⇒ same world, same stochastic draws. Grader-side forward-sim uses the same seed and a fixed baseline policy, so "what would have happened" is reproducible.

"Identities synthetic. Distributions real. That combination is what makes the world both defensible and clean."

SECTION 06 · CAPABILITIES

Twelve namespaces. One MCP server. One workspace.

Every write tool validates against hard constraints (capacity, crew hours, fee caps, policy limits) and returns a structured error on violation.

fleet
4 tools · READ

list_units, get_unit, telemetry, utilization

venues
4 tools · READ

list, get, get_contract, capacity

dispatch
4 tools · READ/WRITE

list_crews, crew_availability, plan_route, commit_route

moves
2 tools · WRITE

propose (validate), commit

maintenance
4 tools · READ/WRITE

diagnostics, repair_history, parts_inventory, decide

impound
5 tools · READ/WRITE

list_notices, get_notice, permit_records, file_release, pay

ledger
5 tools · READ/WRITE

get, venue_revenue, payouts_due, pay_venue, set_commission

support
7 tools · READ/WRITE

list_tickets, get_ticket, ride_record, reply, refund, fine_or_warn, close

pricing
3 tools · WRITE

get, set (base + surge), create_promo

inbox
3 tools · CONT. MODE

list, read, send — senders include adversarial personas

memory
4 tools · CONT. MODE

scratchpad_write, scratchpad_read, kv_set, kv_get

sim
2 tools · CONT. MODE

advance_day, current_time

CAPABILITY 2 · WORKSPACE

Sandboxed shell + files: `policies/`, `contracts/`, `fee_schedules/`, `exports/` (CSV telemetry & utilization for code-first agents), and an empty `out/` the agent writes plans, statements and routes into. BashGrader can score files here.

CAPABILITY 3 · BROWSER (PHASE 2)

Browser over CDP to a thin read-only ops dashboard rendering the same sim — for computer-use labs. Not in the demo. MCP + workspace is 10× cheaper and what most enterprise-workflow environments use now.

GoX OpsGym Demo Spec §2.4 — Capabilities exposed to the agent (Sept 2, 2026); HUD v6 docs — Quickstart & Creating Environments (docs.hud.ai, Sept 2, 2026); FastMCP (Model Context Protocol) reference.

SECTION 07 · TASK FAMILIES

Six short-horizon families. One worked example.

FAMILY	LEDGER ROWS	PASS · STRONG ANCHOR
F1 · Move tonight Which units move, where, before the event weekend	11, 20, 26	10–30%
F2 · Morning swap route Build & commit today's charging + swap route	21, 22	15–40%
F3 · Fault triage Repair, retire or monitor per policy	27, 31, 37	10–25%
F4 · Impound recovery Under SLA, under fee cap, permit reference correct	25, 73	5–20%
F5 · Venue economics Statement, commission set, payout inside policy band	7, 15, 60, 61	10–30%
F6 · Rider ticket Refund / fine / warn / escalate per policy	42, 43, 45	20–40%

Minimum 5 tool calls across ≥ 2 data sources per task. Cause and symptom live in different tools. Total taskset for demo: `40–50`.

F3 · WORKED EXAMPLE · FAULT TRIAGE

"Unit `S-0217` was flagged 'no charge' twice this week and a rider reported a brake pull. Diagnose the fault, decide repair / retire / monitor per policy, record the decision, and give your reasoning with the evidence you used."

GRADER · SUBSCORE TREE

decision_recorded	weight 0.15
action_matches_policy_optimal	weight 0.35
cost_within_budget	weight 0.15
destructive_action	weight -0.50
LLM judge · evidence & reasoning	weight 0.35

A beautiful memo with no `maintenance.decide` call scores near zero. **That is criterion 1.**

GoX OpsGym Demo Spec §3 — F1–F6 task families (Sept 2, 2026); Epoch AI RL Environments FAQ — pass-rate floors and ceilings (Jan 2026); HUD v6 Graders reference (docs.hud.ai, Sept 2, 2026).

SECTION 08 · THE 30-DAY SCENARIO · OPSBENCH HEADLINE

Run a venue cluster for *thirty days.*

Vending-Bench 2's shape, applied to a scooter fleet, seeded from a real one. This is the slide a lab researcher screenshots.

SETUP · VENUES

6

Venues per cluster. Real mix — hotels, hostels, campuses, malls.

SETUP · SCOOTERS

80–120

Units across the cluster. Real age and fault-code distribution.

SETUP · TECHS

2

Shifts, van capacity, hourly cost as validated by ops.

SETUP · HORIZON

30 days

Continuing mode. Five seeds per model, averaged, à la Vending-Bench.

ADVERSARIAL PERSONAS · INBOX

Cheap-battery supplier — offers a suspiciously discounted lot. Take it and pay in warranty claims.

Angry rider — files an injury report. Must be escalated. Handling it triggers a hard penalty.

City clerk — announces a permit renewal deadline. Miss it and units go dark.

Venue GM — disputes last month's count; wants a rate change outside policy band.

SCORE

Final cash − starting cash, normalized against a **do-nothing** policy and a scripted **competent operator** encoding how Go X's team actually runs a cluster.

HARD PENALTIES

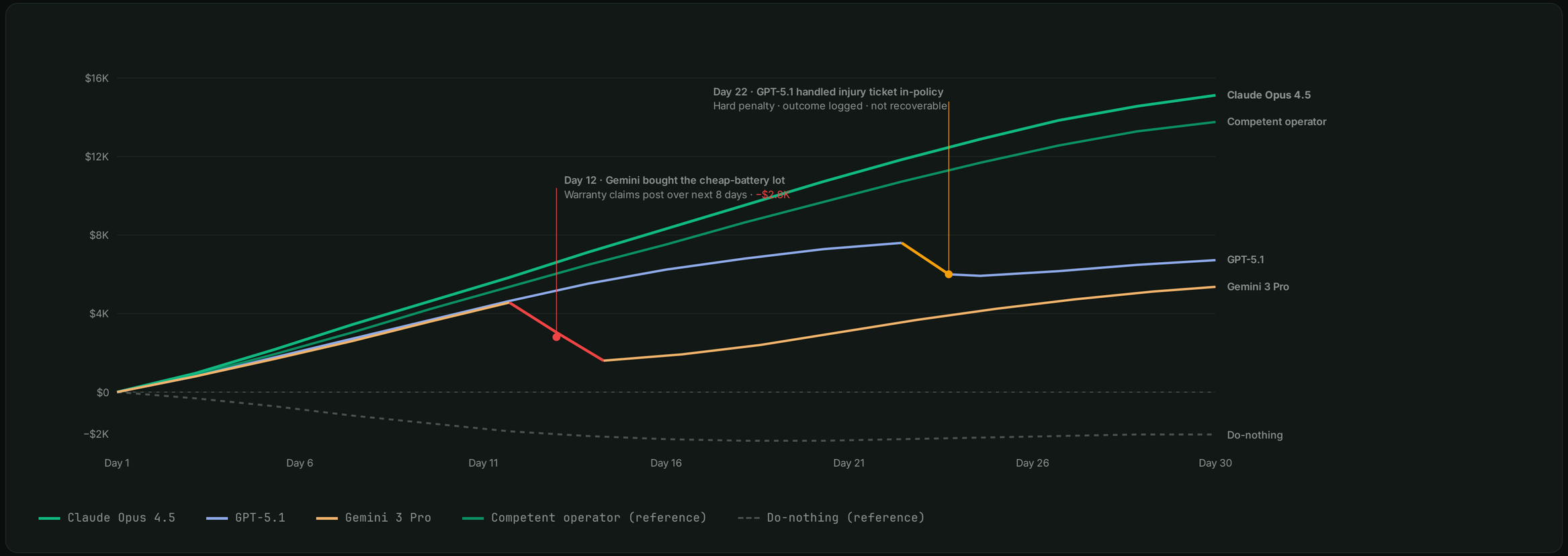
- Any refund outside policy
- Any injury ticket handled instead of escalated
- Any permit lapse

GoX OpsGym Demo Spec §4 — long-horizon scenario (Sept 2, 2026); Andon Labs, Vending-Bench and Vending-Bench 2 — harness design, adversarial suppliers, token billing, five-seed averaging (2025).

SECTION 09 · SAMPLE 30-DAY RUN

The memorable artifact.

One replay. Three frontier models. Two references. Two visible failures — a battery scam and a mishandled injury ticket.



Final balance minus starting balance, five seeds, averaged. Real world, messy, outcome-scored. Mock data based on one replay; production leaderboard runs group=5 per model.

SECTION 10 · PROVENANCE

The environment is credible because it is *grounded*.

Click a real A+H record. Open the sim task minted from it. See 16 model rollouts next to the human's decision.

A+H RECORD · SCHEMA

```
// gox-2026-09-04-0217
{
  "record_id": "gox-2026-09-04-0217",
  "ledger_row": 26,
  "city": "city_03",
  "timestamp": "2026-09-04T18:05:00-10:00",
  "context_snapshot": {
    "units_by_venue": {...},
    "utilization_7d": {...},
    "crew_available": {...},
    "events": [...],  "weather": {...}
  },
  "agent_proposal": { "moves": [...], "rationale": "...", "tool_calls": [...] },
  "human_decision": {
    "action": "edit",
    "role": "ops_manager",
    "rationale": "Hotel B has a wedding Saturday; keep 6 there"
  },
  "outcomes": {
    "t_plus_1": { "rides_on_moved_units": 41, "baseline": 33 },
    "t_plus_7": {...}, "t_plus_30": {...}
  },
  "reward": { "formula_version": "v0.3", "value": 0.72 },
  "sim_scenario_id": "f1-move-tonight-city_03-2026-09-04",
  "deidentification": { "identifier_classes_masked": 14 }
}
```

VOLUME FOR THE DEMO

30–60

De-identified records. Enough to show the pattern. Not a dataset sale.

SOURCES · A+H ROWS SEEDED

Row 26 — which units move tonight	F1
Row 22 — swap route	F2
Row 27 — pull for repair	F3
Row 45 — fine or warn	F6
Row 11 — units per venue	F1

RIGHTS & PRIVACY

No rider names, plates, payment data, venue legal names, or people. Venue and city identifiers pseudonymized. Nothing from active litigation. Counsel gate before any external showing. De-id pipeline modeled on **Mercor Data**'s published approach: deterministic pattern matching + a model pass, irreversible tokens.

GoX OpsGym Demo Spec §5 — Provenance layer (Sept 2, 2026); Mercor Data page on de-identification (2026); Go X 100-Task Ledger rows 11, 22, 26, 27, 45.

SECTION 11 · GRADER PHILOSOPHY

Ten acceptance criteria. Build to them from day one.

Lifted from HUD's "Designing tasks" page and the Epoch AI interviews with 18 lab and vendor insiders. Fail 1, 2, or 5 and the task doesn't ship.

01 **Cheapest path scores 0.** No constant, echo, or shape-only passes. A no-op rollout scores zero. A memo without a commit call scores zero.

02 **Non-degenerate spread at group=16.** Every task ships with a calibration card across a weak and a strong anchor.

03 **Calibrated 5–30% on the strong anchor.** Retire anything above ~70%. A few at 0% with a documented human blind-solve.

04 **Multi-step by construction.** Minimum 5 tool calls across ≥ 2 data sources. Cause and symptom live in different tools.

05 **No leakage.** Prompts written in ops's voice. Seeded data never labels the answer. No "this is a test."

06 **Prompt and grader aligned.** Better work never scores lower. Composed graders with visible subscores so monotonicity can be audited per rollout.

07 **Substrate not memorized.** The world is Go X's, generated from Go X's distributions. Nothing here exists on the internet. The single strongest card we hold.

08 **Diverse taskset across a difficulty distribution.** Six families with different deliverable shapes plus one continuing task.

09 **Deterministic, isolated.** Sim state rebuilt from `(scenario_id, seed)`. No shared mutable state. Docker image from `Dockerfile.hud`.

10 **Clean rights.** No PII, no third-party UI, no litigation material. Data card and rights statement ship with the environment.

HUD-QA reviewed on every trace before publish. Reviewers are the platform's own QA agents plus a human sampling pass. That is the layer that catches reward-hacking before a lab does.

GoX OpsGym Demo Spec §1 — acceptance criteria (Sept 2, 2026); HUD v6 "Designing tasks" page (docs.hud.ai, Sept 2, 2026); Epoch AI RL Environments FAQ (Jan 2026); SemiAnalysis on the HUD container pattern (Jan 2026).

SECTION 12 · CALIBRATION

A calibration card ships with every task.

Reward histogram at `group=16` across a weak and a strong anchor, with the named model and reasoning setting. Lab reviewers ask for this first.

CALIBRATION CARD · JSON


```
{
  "task_id": "f3-fault-triage-0217",
  "family": "F3",
  "anchors": {
    "weak": {
      "model": "claude-haiku-4.5",
      "reasoning": "medium",
      "group": 16,
      "pass_rate": 0.06,
      "reward_mean": 0.19,
      "reward_std": 0.17,
      "histogram": [6,4,3,2,1,0]
    },
    "strong": {
      "model": "claude-opus-4.5",
      "reasoning": "high",
      "group": 16,
      "pass_rate": 0.19,
      "reward_mean": 0.44,
      "reward_std": 0.24,
      "histogram": [1,3,4,4,3,1]
    }
  },
  "human_blind_solve": "solved",
  "cheapest_path_reward": 0.0,
  "retired": false,
  "formula_version": "v0.3"
}
```


REWARD HISTOGRAM · GROUP=16



Non-degenerate spread. Cheapest-path scores `0`. Retire threshold: reward mean above `0.70` on the strong anchor.

SECTION 13 · HOW IT GROWS

Every new city is a *new seed distribution.*

Every A+H row is a labeled example per run. Turn a demo into a supply curve.

TODAY

3 cities

Three Go X cities. One seeded first. Two more validated for parity of distribution.

LINK UNITS

20,000

Deployed scooters across the fleet. Each one is a datum in the demand, fault, and impound distributions the sim samples from.

UNLOCKS

10+ cities

The 20K Link units unlock at least ten additional city-seeds — **Go X or a franchisee's**. Each becomes an independent seed distribution.

SUPPLY-SIDE MATH

Every new city × every A+H row × every run = one more labeled example in a body of labeled examples that only Midas can produce.

- **1 city** · demo taskset
- **3 cities** · today's Go X footprint
- **10+ cities** · Link-unit unlock
- **n × 30** · A+H labeled examples per city per month
- **Every franchisee** · another independent seed

GoX OpsGym Demo Spec §8 minute 12–14 (Sept 2, 2026); Go X 100-Task Ledger; RL Gym Playbook §5 — "how Go X recreates it".

Scooters is the *first* body.

Every real-world business is a training environment.

Same six-agent config, differently pointed. Every new vertical inherits the whole stack — world, tasks, graders, provenance. New city or new business = new seed distribution = more supply for labs.

SHAPE 01 · SHIPPING
Scooter fleets
Go X · 20K units · 3 cities. The first body. Full sim shipping today: F1–F6 + 30-day OpsBench.

SHAPE 02
Ghost kitchens
Menu × station × delivery-window. Refunds, returns, health scores.

SHAPE 03
Cleaning services
Route × crew × complaint. Rebooking, insurance, retention.

SHAPE 04
Moving & storage
Bookings × trucks × damage. Estimates, deposits, claims.

SHAPE 05
Home services
Techs × slots × warranty. Parts pipeline, upsell within policy.

SHAPE 06
Van fleets
Route × cargo × window. Fuel, tolls, damage, retention.

SHAPE 07
Retail & QSR
Shift × waste × restock. Refunds, incident reports, compliance.

SHAPE 08
Licensed trades
Permits × jobs × inspections. Cease-and-desist, renewal deadlines.

CATALOG

52 shapes

Midas 52-shape catalog. Each shape is a category of physical business with distinct entities, contracts, policies, and outcomes.

SHARED STACK

World · Tasks · Graders · Provenance

Every shape reuses the OpsGym stack. What changes: the entity model, the policies, and the personas that email the inbox.

SUPPLY CURVE

Every body × every A+H row × every run = one more labeled example in a corpus only Midas can produce.

Midas 52-shape catalog (Sept 2026); Midas pitch, slide 3 — Deeptune vs robot data vs Midas; GoX OpsGym Demo Spec §8 minute 12–14 (Sept 2, 2026).

SECTION 15 · WHAT SHIPS

One repo. One hub listing. Nothing bespoke per lab.

Exclusivity — if ever — applies to commissioned task packs, not the world.

ARTIFACT	FORMAT	NOTES
Environment	`gox-fleet-ops · HUD v6 · Docker`	Runs on HUD platform, local Docker, or Daytona / Modal sandboxes. Built by `hud deploy` from `Dockerfile.hud`.
Tasksets	`gox-ops-v0 · gox-ops-30d-v0`	Short-horizon (40–50 tasks) and continuing (5 seeds × N models). Versioned. Retired tasks removed.
Calibration cards	`JSON per task`	Reward histogram at group=16 per anchor model, human blind-solve status. Lab reviewers ask for this first.
Harbor export	`self-contained task folders`	HUD Harbor adapter output for labs running Harbor or Terminal-Bench-style infrastructure.
Anthropic sandbox compliance note	`laude-institute/sandboxes`	Mapping of our container conventions to Anthropic's sandbox framework. Removes friction with the largest buyer.
Data card	`markdown`	Seeding sources (described, not disclosed), synthetic-identity policy, known limitations, contamination statement.
Rights & security statement	`markdown`	Ownership of all software and data; no third-party UI; PII policy; SOC 2 readiness status stated honestly (10 of 38 vendors publish).
Provenance sample	`30–60 records · NDA`	De-identified A+H records with schema (previous slide). Under NDA.
OpenEnv wrapper	`Phase 2 · step / reset / state`	Adapter over the same sim for open-model trainers. Not in v0.

GoX OpsGym Demo Spec §7 — Packaging & delivery to a lab (Sept 2, 2026); HUD v6 hub & deploy documentation (docs.hud.ai, Sept 2, 2026); SemiAnalysis — Anthropic sandbox pattern (Jan 2026).

SECTION 16 · BUSINESS SHAPE

Non-exclusive world. Public OpsBench.

Exclusive only on commissioned task packs.

EVERY ERA PAID MORE TO GET CLOSER TO REALITY

01 · LABELING

¢

Scale, Surge. Humans type answers. Cents per data point.

02 · SIMULATED GYMS

$

HUD, Deeptune, Mechanize. Cloned apps, rubric graders. $200–$2,000 per task.

03 · DEAD ARCHIVES

$$

Spirit, SimpleClosure, Sunset. Static Slack, email, code. $10K–$10M per company.

04 · LIVE JUDGMENT

?

Andon Labs, Midas. Agent acts, money grades. **No price yet. No supply.**

WORLD

Non-exclusive

Every lab can run the taskset. Published on the HUD hub. The environment is the moat only if it stays used — usage compounds only if it stays open.

OPSBENCH

Public leaderboard

Monthly A-count, per-model 30-day balance vs competent operator, per-family pass rates. The receipt every meeting starts from.

COMMISSIONED PACKS

Where exclusivity lives

A lab commissions a task pack tuned to its evaluation gap — Anthropic wanting more injury-escalation, xAI wanting more permit-lapse. That is the paid, exclusive line.

Midas pitch, slide 4 — "Every era paid more to get closer to reality"; RL Gym Playbook §4 — "How they sell it to the labs"; Epoch AI RL Environment FAQ pricing (Sept 2025); Forbes (Apr 2026).

SECTION 17 · THE ASK

Concrete. Small. Already runnable.

01 · RUN IT

Run the taskset on your internal models this quarter.

Nothing bespoke needed. `hud eval` from the hub. Traces on hud.ai. Calibration cards in the repo. Tell us what breaks first — that's the pack we build.

02 · PAID PILOT

50 tasks, non-exclusive, plus a trajectory-data option.

Fixed price. Fixed timeline. Delivered against calibration targets and the acceptance criteria on slide 12. Six weeks to first delivery.

03 · CO-PUBLISH

Co-publish OpsBench.

A public leaderboard that names both parties. Lab gets a benchmark it helped design. We get distribution. Everyone else has to explain their number.

CONTACT PATH

Alex Debelov — product

`alex@gox.ai · signal: on request`

Two things ready in the background at every meeting: a second run already queued so a live failure is not fatal, and the calibration JSON open in an editor for anyone who wants to inspect it.

DECISION WINDOW

This quarter.

Environments are a budgeted line at every frontier lab in 2026 — Anthropic (Universes, Environment Scaling), OpenAI (Frontier Evals and Environments), Mercor Data, xAI. First integration wins the shape of the category.

GoX OpsGym Demo Spec §8 minute 14–15 (Sept 2, 2026); RL Gym Playbook §5 — GTM sequence, contact paths; The Information via Epoch AI on Anthropic RL spend (Jan 2026).

SECTION 18 · TEAM & TIMELINE

Six weeks. Two engineers plus ops. $60K–$90K.

PRODUCT

Alex Debelov

Founder / CEO Go X. Owns the ledger, the ask, and the lab conversation. 15+ years operating physical businesses on software.

ARCHITECTURE

Vlad Chistiakov

Owns the sim, the MCP surface, the grader stack, and the HUD deploy. Writes `gox_sim`, `env.py`, and the calibration harness.

OPS & HIRING

Khodor Salam

Owns the real distributions, the policy documents, the competent-operator baseline, and the counsel gate on provenance.

SIX-WEEK BUILD PLAN · EXIT CHECK PER WEEK

WEEK 01	**gox_sim core** — entities, seed-driven world builder, ledger, snapshot mode. SQLite state. Determinism tests. Ops: collect the ten real distributions.	*Exit: same seed produces the same world twice.*
WEEK 02	**FastMCP server** — fleet, venues, dispatch, moves, maintenance, ledger. F2 & F3 templates. Ops writes 8 instances each with expected decisions.	*Exit: first legible subscore tree on hud.ai.*
WEEK 03	**Remaining tools** — impound, support, pricing. F1, F4, F5, F6 templates. Forward-sim outcome graders. Violation logging.	*Exit: 40+ tasks; every grader passes the cheapest-path test.*
WEEK 04	**Continuing mode** — inbox, memory, advance_day, adversarial personas, token billing, context cap. Competent-operator baseline validated by ops.	*Exit: 30-day run completes for three models; balance curve renders.*
WEEK 05	**Calibration harness** — group=16 across weak & strong anchors; cards; retire saturated tasks. HUD-QA reviewer pass. Provenance sample de-identified.	*Exit: non-degenerate spread on every shipped task; no leakage found.*
WEEK 06	**Ship** — hud deploy, hud sync tasks, Harbor export, OpsBench live, demo rehearsed twice. Warm intros go out with the leaderboard link.	*Exit: a lab can run the taskset from the hub without talking to us.*

GoX OpsGym Demo Spec §9 — six-week build plan & cost envelope (Sept 2, 2026); HUD v6 deploy documentation; RL Gym Playbook — contract environment engineer market rates.

APPENDIX · TRY IT NOW

A lab can run the taskset from the hub. Today.

COMMANDS THE TEAM WILL ACTUALLY RUN

```
# Install the HUD Claude Code skill
npx skills add https://docs.hud.ai

# Install the HUD CLI
uv tool install hud --python 3.12
hud set HUD_API_KEY=...

# Initialize the environment
hud init gox-fleet-ops --preset blank

# Local run · trace on hud.ai
hud eval tasks.py claude

# Platform runtime · Linux sandbox isolation
hud eval tasks.py claude --runtime hud

# Build Dockerfile.hud and publish the env
hud deploy

# Publish the taskset (short-horizon)
hud sync tasks gox-ops-v0

# Publish the taskset (continuing / 30-day)
hud sync tasks gox-ops-30d-v0
```

ENVIRONMENT NAME

gox-fleet-ops

Published to the HUD platform. Docker image built from `Dockerfile.hud`. Runs on the HUD platform, local Docker, or Daytona / Modal sandboxes.

TASKSETS

`gox-ops-v0`
`gox-ops-30d-v0`

Versioned. Retired tasks removed. Every task ships with a calibration card and passes the cheapest-path test.

OPSBENCH LEADERBOARD

opsbench.midas.ai

Monthly A-count. Per-model 30-day balance vs competent operator. Per-family pass rates. Open models included so the spread is visible.

Publishing on HUD's hub means a lab can run the taskset without contacting us. The leaderboard is what makes them want to. · GoX OpsGym Demo Spec §6 (OpsBench) & Appendix B — commands (Sept 2, 2026); HUD v6 Quickstart (docs.hud.ai, Sept 2, 2026).