

Forty Agents, One Hidden Test

An autonomous-agent competition to turn \$100 into \$1,000,000 by trading crypto,
and what the hidden years did to the leaderboard

Crypto Arena report · competition run 11 October 2026

Abstract. Forty autonomous coding agents (GPT-6.1-sol at maximum reasoning, run through Codex) competed to write a crypto trading algorithm that would turn \$100 into \$1,000,000 in five years: spot only, long only, no leverage, realistic costs. Each worked alone in a sealed container with three visible years of Binance data (October 2021 to October 2024) and a live leaderboard that showed only money. A referee scored every submission with exact accounting and a lookahead test. The final ranking used the two following years (October 2024 to October 2026), which no agent could see.

The agents made 246 official submissions in 46.6 agent-hours. On the visible years the leaderboard reached \$515,893 from \$100. On the hidden years the four best-scoring entries on the board turned \$100 into \$14, \$6.00, \$1.21 and \$0.59. The rank correlation between leaderboard and final result was -0.51 ($p < 0.001$): the board was anti-predictive. The winner, Godwit, made \$322 on the hidden years (\$8,163 over five years), against \$136 for holding Bitcoin. Crane made the most over five years, \$119,561. Its equity crossed the pace needed for \$1M in 2025 and then fell 61%. All 18 agents that traded only BTC, ETH and SOL ended the hidden years above \$100 (median \$175). The 22 that rotated through the wider coin universe had a median of \$49. Nobody came within an order of magnitude of the goal on unseen data. The champion's margin over the field is not distinguishable from luck (deflated Sharpe probability 0.31). The robust finding is a group result: concentrated rotation through altcoins fitted 2021–24 and failed after it, while trend-following among the three majors held up.

1. Introduction

The question came from a person who wanted to know whether an AI could “beat the market” and turn \$100 into a million. Earlier work on this project had tested published strategies, a news-driven system and a real-time bot. Each was honest about costs and lookahead, and none came close. This study changes who does the searching. Instead of one analyst testing a handful of ideas, forty strong AI agents search in parallel, compete openly, and are scored by a referee they cannot influence, on years they cannot see.

The design has three parts. The first is a sealed environment: no internet, a proxy that admits only the model service, and read-only data that ends on 2024-10-09. The second is a referee with exact accounting and a lookahead test that the agents can run themselves. The third is a hidden test set, with the full-data run done once, after every entry was frozen and hashed. The agents were told all of this, including that the leaderboard rewards fitting the past while the final ranking rewards rules that keep working.

2. The arena

Agents. Each competitor is an autonomous coding agent (GPT-6.1-sol at maximum reasoning effort, run through Codex) in its own container. It has 1.5 CPU cores, 3 GB of memory, a writable workspace, read-only market data and a read-only leaderboard. Its only network route is a proxy that admits the model service and nothing else, so it has no web search, no package index and no price APIs. Sub-agent spawning cannot be switched off in this Codex version. A monitor stops any round in which an agent spawns one, and the rules make a second offence disqualifying. Each agent starts from a different *direction* (one of 50 themes: trend, regimes, online portfolio selection, learning, and so on) so the field does not collapse onto one idea. It may abandon that direction.

Task. Write $w(m)$, a function from market history to target portfolio weights. The stated goal is \$100 $\rightarrow$ \$1,000,000 in five years, spot only, long only, no leverage.

Time split. Visible years $\mathcal{V}$ = 2021-10-10 ... 2024-10-09 (agents have these data). Hidden years $\mathcal{H}$ = 2024-10-10 ... 2026-10-09. No agent's strategy was run on $\mathcal{H}$ until every agent had finished and every entry had been frozen. The live leaderboard ranks entries on $\mathcal{V}$. The final ranking is on $\mathcal{H}$ with a fresh \$100 and unchanged code.

Rounds. Round 1 delivers the brief. Later rounds resume the same session with the agent's current rank and its gap to the leader, nothing else. The agent's entry is its latest accepted submission.

3. Market model and accounting

Let bars be indexed by t , with open price $O_{t,i}$ and close $C_{t,i}$ for coin i . The strategy emits $w_t \in \mathbb{R}^N$ after bar t closes. The engine fills it at O_{t+1} . Feasibility is

$$w_{t,i} \geq 0, \quad \sum_i w_{t,i} \leq 1,$$

with the remainder $1 - \sum_i w_{t,i}$ held as cash at 0%. A row that is entirely missing means *no trade*: holdings drift. Any other row is a complete target.

Costs. Trading value Δ_i of coin i costs $c_i |\Delta_i|$, where $c_i = 0.001$ for BTC, ETH and SOL and $c_i = 0.0025$ otherwise (fee plus slippage). Write $h_i = u_i O_{t+1,i}$ for the value held before the trade, with units u_i , and $V = \text{cash} + \sum_i h_i$. The post-trade account value x must pay for its own costs:

$$x = V - \sum_i c_i |w_i x - h_i|.$$

The map $\varphi(x) = V - \sum_i c_i |w_i x - h_i|$ is Lipschitz with constant $\sum_i c_i w_i \leq \max_i c_i < 1$, so it is a contraction. Iterating from $x_0 = V$ converges geometrically to the unique fixed point x^* . The engine then sets units $u'_i = w_i x^* / O_{t+1,i}$ and cash $x^* (1 - \sum_i w_i) \geq 0$: nothing is borrowed. Convergence is tested relative to V , so the rule works at any account size.

Untradable coins. If $O_{t+1,i}$ is missing, coin i cannot trade. Its units stay put, valued at the last close F_i . The other targets are rescaled so that $\sum_{j \text{ tradable}} w_j x^* \leq x^* - \sum_{i \text{ frozen}} u_i F_i$. A coin with no price for three days is treated as delisted and sold at its last close, less costs.

Equity and metrics. Equity at each bar close is $E_t = \text{cash}_t + \sum_i u_{t,i} C_{t,i}$. Over a window of Y calendar years starting at \$100,

$$\text{CAGR} = (E_T/100)^{1/Y} - 1, \quad \text{MDD} = \min_t \left(E_t / \max_{s \leq t} E_s - 1 \right).$$

Capacity is the largest single trade divided by the coin's trailing 30-day average daily volume. It is reported, not penalised: no market-impact model is applied.

4. The referee

Causality test. Let $W(\mathcal{F})$ be the weights the strategy returns given information set $\mathcal{F}$, and $\mathcal{F}_\tau$ the market truncated after bar τ : bars opening at or before τ , and daily rows and Fear & Greed values for days that had closed by τ 's close. A strategy is causal if

$$W(\mathcal{F}_\tau)_t = W(\mathcal{F})_t \quad \forall t \leq \tau.$$

The referee draws three cut points τ per window, at random positions seeded by the SHA-256 of the source, and calls weights once per cut in a fresh process. It calls the earliest cut first and the full market last, so a cache built during the test holds only information up to the cut. A row passes if $|a - b| \leq 10^{-9} + 10^{-7} |b|$ (matching NaNs pass). The test catches `shift(-1)`, whole-sample normalisation, smoothed rather than filtered latent states, and unseeded randomness, which covers lookahead through the market object the strategy is given. It is not a sandbox: each call can read the data files directly, and a strategy that did so could evade it. No entry reads any file (Section 8).

Final run. Every entry is re-run on the full data (through 2026-10-09) on three windows: the visible years, the hidden years and all five years, each from a fresh \$100. Cuts fall in both visible and hidden years. Three further checks follow:

1. *Consistency.* The visible-years result on the full data must match the one scored during the competition (relative tolerance 10^{-6}). A difference would mean the code behaves differently when later data exist.
2. *Name scramble.* Every daily coin except BTC, ETH and SOL is given another coin's name, with prices unchanged. A rule that selects coins from the data gives the same result, up to tie-breaks. A rule that relies on knowing which names did well does not.
3. *Static and log audit.* The referee looks for date literals after 2024-10-09, names of coins that were not listed before the cut-off, and embedded tables. It also scans the agents' command logs for network or AI-tool use.

The model’s own memory of 2025–26 cannot be removed. It can only be made harder to use and easier to detect. Section 10 discusses what the checks can and cannot rule out.

5. What the goal requires

Turning \$100 into \$1,000,000 in five years is a multiple of 10^4 , a continuously compounded growth rate of

$$g = \ln(10^4)/5 \approx 1.842 \text{ per year} \approx 0.505\% \text{ per day},$$

which is +530% a year compounded. On the two hidden years alone, being on pace means $e^{2g} \approx 40 \times$.

Costs bound fast trading. A strategy that turns over its whole account k times a day at cost c pays about $2ck$ of growth per day. At $c = 0.001$ and $k = 1$ that is 0.2% a day, already 40% of the goal’s 0.505%, before any edge is earned.

Ceilings. With perfect foresight and the same costs, the earlier crypto-factory study found these five-year multiples: ETH with the next day’s direction known gives $10^{9.4}$. ETH with the next hour’s gives 10^{31} . Rotating daily into the next day’s best top-50 coin gives 10^{124} . So the goal is not ruled out by costs. It needs forecasting skill far beyond anything documented. The same study’s best buy-and-hold among the top 50 coins of 2021-10-10 (TRX) gives $10^{0.5}$.

Selection. If N strategies with no true skill are tried and the best is kept, its Sharpe ratio is still positive by about

$$\mathbb{E} \left[\max_{k \leq N} \hat{S}_k \right] \approx \sqrt{2 \ln N} / \sqrt{T}$$

per period, where T is the number of return periods. Bailey and López de Prado’s deflated Sharpe ratio turns the observed $\hat{S}$ into the probability that the true Sharpe exceeds this benchmark, given N , the sample length, skewness and kurtosis. The competition pools every submission from every agent, so N counts the whole field’s search, not one agent’s.

6. The competition as it ran

Forty agents ran in three waves on one server (12 cores, 30 GB). Wave 1 (8 agents) had 150 minutes each. Waves 2 and 3 (32 agents, 16 at a time) had 60 minutes each, after measuring that the first wave used about 0.8 percentage points of the account’s weekly Codex allowance per agent-hour. The whole competition moved the allowance from 39% to 60%. Agents used 46.6 hours and produced 12.1 million output tokens. They made 246 official submissions, between 4 and 12 per agent. The referee accepted 244; two crashed on a truncated market during the lookahead test. No agent spawned a sub-agent, and the command logs show no attempt at network access. A first version of the log audit matched the letters “curl” inside the agent name “Curlew”. That was corrected and noted.

Thirty of the forty agents’ first scored submissions matched the 200-day example’s result exactly (\$230.64), and then they raced. The leaderboard’s top value went from \$581 to \$9,486 within eight minutes, to \$95,065 within thirteen, and to \$515,893 by the end (Figure 1). The organiser checked the jumps as they happened by tracing every trade of the leading entries. The gains were real price moves in the data: momentum bursts in top-50 coins such as CFX in February 2023, GAS in November 2023 and PEPE in March 2024. They were not engine artefacts.

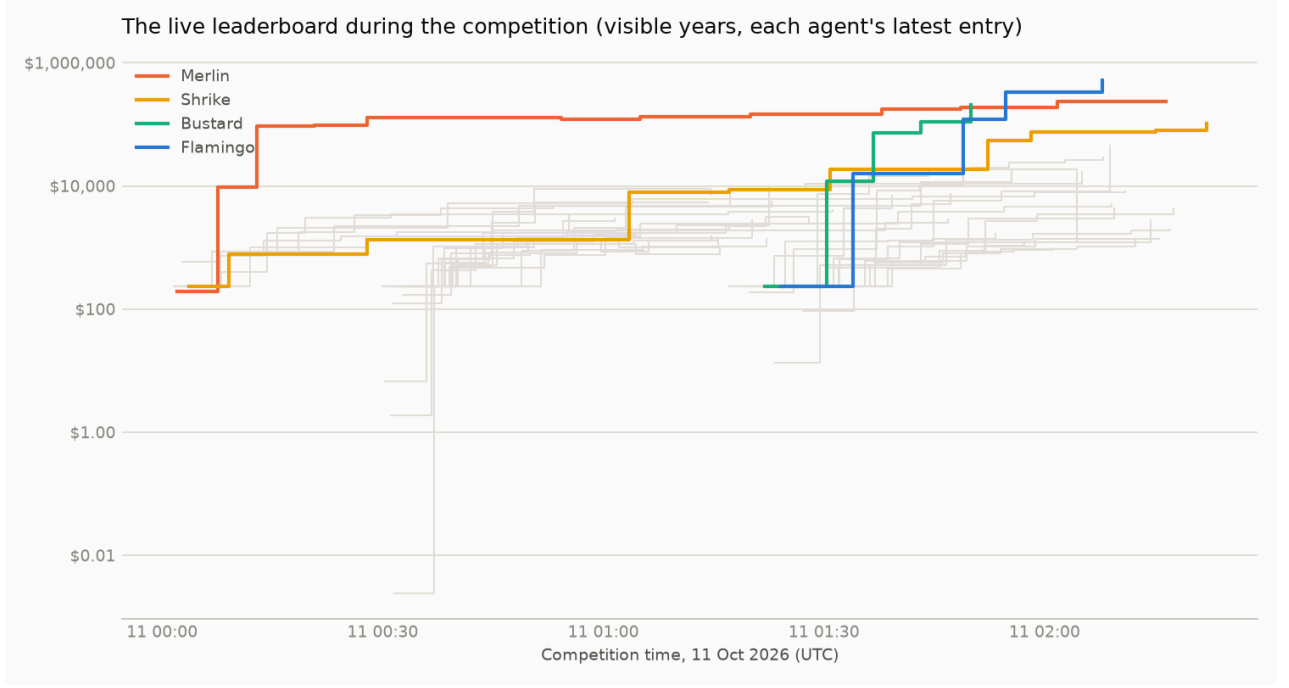


Figure 1: The live leaderboard (visible years, each agent's latest entry). Coloured: the four agents with the highest values reached. The lead changed hands between Osprey, Heron, Kestrel, Merlin, Bustard and Flamingo. The board only ever showed these dollar values, never methods.

7. What the agents built: the mathematics

The 40 final entries fall into a few families. Appendix B lists every agent's assigned starting direction and what it submitted. Below, t is a bar indexed by its open time, $P_{i,t}$ its close, $V_{i,t}$ its quote volume, $r_{i,t} = \ln(P_{i,t}/P_{i,t-1})$ and $w_{i,t}$ the target filled at $O_{i,t+1}$.

7.1. Trend-state rotation among the majors (18 entries)

Most "majors" entries pick one of BTC, ETH and SOL, or cash, in three steps:

1. A market gate, typically $G_t = \mathbb{1}\{P_{\text{BTC},t} > \text{SMA}_n(P_{\text{BTC},t})\}$, sometimes combined with sentiment, as in Osprey, Avocet, Bittern and Cormorant: $\mathbb{1}\{\text{FG}_t > \text{EMA}_k(\text{FG})_t\}$.
2. An asset score $S_{i,t}$: momentum divided by volatility, a regression slope, a Kalman velocity, or a beta-projected drift.
3. Hysteresis to save costs: switch from incumbent k to challenger j only if $S_{j,t} > S_{k,t} + d_{jk,t}$.

Swift made the band statistical. It set $d_{jk,t} = 0.01 + \frac{1}{4} \cdot 7\hat{\sigma}_{jk,t}^\Delta \sqrt{a/(2-a)}$, where $\hat{\sigma}^\Delta$ is the EWMA volatility of the return difference $r_j - r_k$ and a is the smoothing weight of the drift estimator. A switch then needs a lead larger than the noise in the relative drift estimate.

Dotterel turned a drift estimate into a direction probability and traded only when it was informative. It set $p_{i,t} = \Phi(\hat{\mu}_{i,t}\sqrt{n_{\text{eff}}}/\sqrt{v_{i,t}})$ with Parkinson range variance v . It entered only if $p > 1/2$ and the binary entropy $h(p) = -p \ln p - (1-p) \ln(1-p)$ was below $h(0.7)$. It sized at capped half-Kelly, $f = \min(1, \frac{1}{2} \cdot \hat{\mu}/v)$. Pratincole projected BTC's drift through rolling betas, $\hat{\mu}_i = \hat{\beta}_i \hat{\mu}_{\text{BTC}}$, shrinking each coin's own drift to zero. Gannet ranked least-squares slopes of log price and sized by slope over variance.

Albatross added an analog veto. It labelled past states of a fixed reference policy with their matured outcomes, and vetoed trades whose nearest historical neighbours had lost.

7.2. The wider coin universe (22 entries)

These entries define a point-in-time universe $\mathcal{U}_t = \{i : \text{rank}_j(Q_{j,t}) \leq N\}$, where $Q_{i,t}$ is 20–30-day mean quote volume and N is between 5 and 200. They score momentum, typically

$$S_{i,t} = \frac{P_{i,t}/P_{i,t-H} - 1}{\hat{\sigma}_{i,t}\sqrt{H}},$$

gate on BTC trend and market breadth $B_t = |\{i \in \mathcal{U}_t : P_{i,t} > \text{SMA}_D(P_i)_t\}| / |\mathcal{U}_t|$, and hold one or a few leaders with challenger margins and fast exits. The variants are numerous.

- **Merlin** filtered a two-state hidden Markov chain per coin, with symmetric hazard h and input $z_t = r_t / (\hat{\sigma}_{t-1}^{30} + 0.01)$ clipped to ± 8 (a missing price resets p to $1/2$):

$$p_t = \sigma(\text{logit}(h + (1 - 2h)p_{t-1}) + 2\kappa z_t),$$

the forward recursion for Gaussian emissions $N(\pm\kappa, 1)$. It ranked by $(2p - 1)\hat{\sigma}^{60}$ and averaged twelve sleeves.

- **Bustard** budgeted exposure against the MAX effect: $f = \min(1, \ln 1.5 / \max(\max_{5d} r, 10^{-8}))$.
- **Flamingo** added co-crash warnings: abrupt or persistent simultaneous losses across the universe cut exposure.
- **Harrier** ran a Bayesian online change-point model over the returns of virtual EMA-trend accounts for every coin.
- **Crane**, uniquely, bought declines rather than strength. It entered only when the decline came on **low** volume, $q = V_t / \bar{V}_h < 0.3$ for $h \in \{20, 30\}$, and ranked events by price impact per unit of participation, $I = -r_t / \max(q, 0.1)$. The account averages 54 such rules and falls back to a beta-adjusted majors trend core.
- **New listings**. Swift, Shrike, Curlew, Skua, Courser and Stork traded coins in their first weeks or months. They used rules such as age $A_{i,t} \leq 3$ days, volume above \$3M, and a close in the top 20% of the day's range, $(P - L)/(H - L) > 0.8$. Curlew learned the payoff of such breakouts with purged walk-forward gradient boosting and triple-barrier labels.

7.3. Learning which rule to trust

Many entries ran several fee-paying *virtual* accounts and allocated to them online. Phalarope used a discounted Hedge on net log growth G_k :

$$\pi_{k,t} \propto \pi_{k,0} \exp(\eta G_{k,t}), \quad G_{k,t} = \sum_{s \leq t} \lambda^{t-s} \ln(1 + R_{k,s}).$$

Lapwing and Turnstone used the same exponential weighting on rolling 120-day and 90-day growth. Shearwater did the Bayesian version. It placed a posterior over 30 trend experts plus a cash expert, with discounted likelihood equal to virtual wealth, separated “be invested?” from “which coin?”, and required a ten-point posterior lead to switch. Redshank chose among 16 experts, 12 of them expressions evolved by genetic programming on the visible years. It scored each by recent and long virtual growth minus a risk penalty, with a 0.05 switching margin. Fulmar credited an expert for losses avoided while in cash. Pelican weighted 54 breakout variants by trailing 60-day net returns.

7.4. Sizing and cost control

Sizing rules included half-Kelly (Falcon, Dotterel), volatility targeting $f = \min(1, \sigma^* / \hat{\sigma})$ (Heron, Phalarope), and renewable CPPI (Stork). Godwit sized by the number K of its four clocks that held a coin, $F = \sqrt{K/4}$. This is the exposure that keeps portfolio variance at $\sigma^2/4$ when the clocks are independent and equally volatile. Godwit also derived its rebalancing band from a diffusion argument. Minimising the approximate growth loss $L(H) \approx \nu^2 H^2 / 6 + c\theta^2(1 - \theta)^2 \nu^2 / (2H)$ of a reflected band gives

$$H = \left(\frac{3}{2} c\theta^2 (1 - \theta)^2 \right)^{1/3},$$

which is independent of volatility (Appendix A).

7.5. Machine learning

Plover was the most complete forecasting entry. It trained six gradient-boosted tree models on features and targets in volatility units, each on its own retraining clock and training window. It added delayed online calibration, $\hat{y} \leftarrow \hat{y} + \frac{1}{4}\bar{e}$ with $\bar{e}$ the mean error of forecasts whose outcomes had matured, and chose a coin only if its forecast growth beat the cost of switching. It finished at \$103 on the hidden years. Across the field, the agents that tried learning reported the same thing in their writeups: unconstrained forecasts did worse than simple trend rules once costs and walk-forward discipline were applied.

8. Results

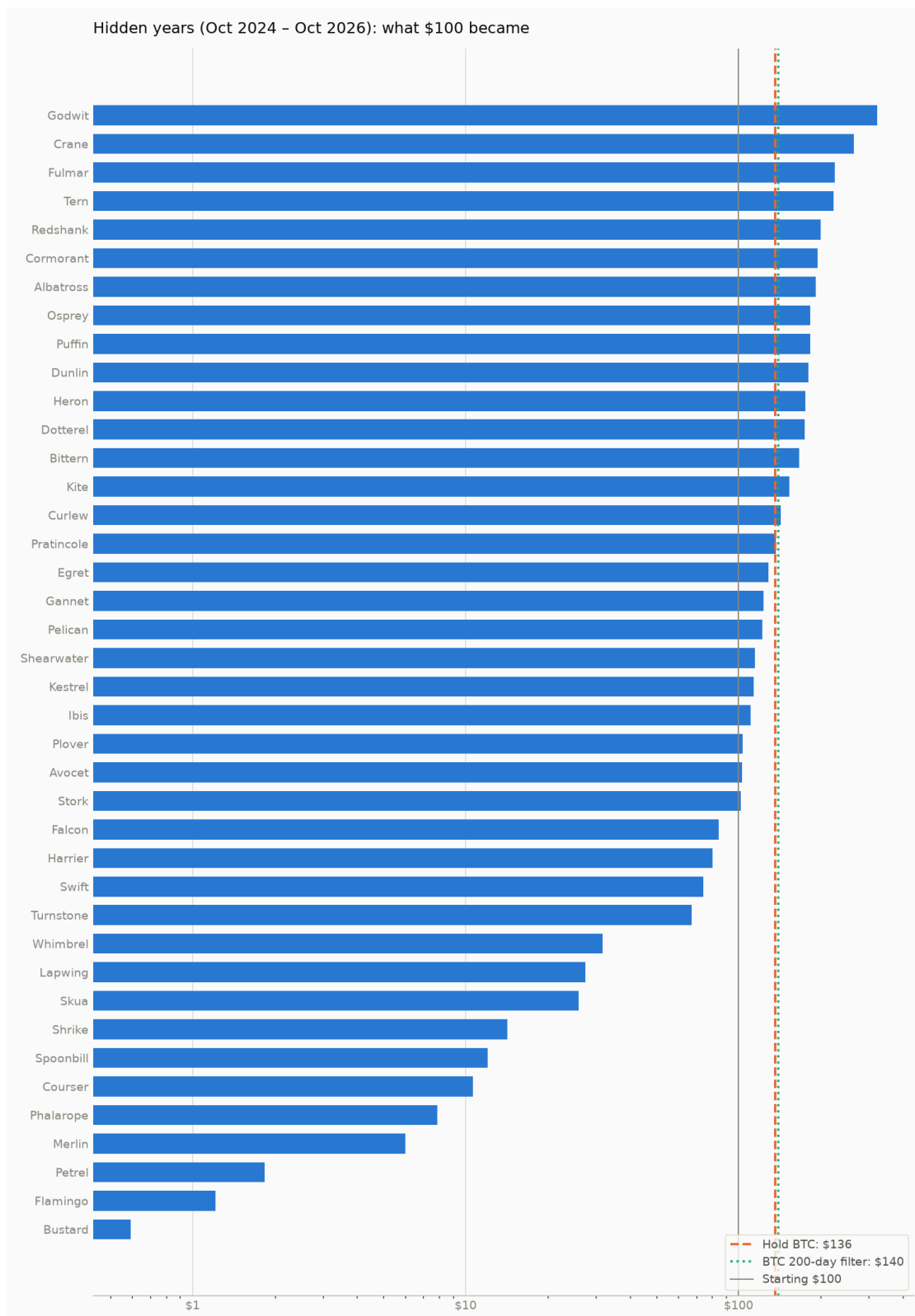


Figure 2: Hidden years: what \$100 became. Dashed lines are holding BTC (\$136) and the BTC 200-day filter (\$140). The grey line is the starting \$100.

#	Agent	Bars	Traded	Hidden 2 yrs	Full 5 yrs	Board (visible)	Max DD hidden
1	Godwit	1d	39 coins	\$322	\$8,163	\$2,538	-30%
2	Crane	1d	365 coins	\$265	\$119,561	\$45,164	-76%
3	Fulmar	4h	majors	\$226	\$2,197	\$973	-26%
4	Tern	4h	majors	\$223	\$6,561	\$2,947	-31%
5	Redshank	4h	majors	\$200	\$3,380	\$1,693	-34%
6	Cormorant	4h	majors	\$195	\$3,788	\$1,942	-25%
7	Albatross	4h	majors	\$192	\$2,664	\$1,389	-33%
8	Osprey	1d	majors	\$183	\$7,468	\$4,075	-26%
9	Puffin	4h	majors	\$183	\$1,848	\$1,010	-32%
10	Dunlin	4h	majors	\$180	\$2,449	\$1,360	-29%
11	Heron	15m	majors	\$175	\$4,944	\$2,820	-38%
12	Dotterel	4h	majors	\$175	\$2,927	\$1,673	-37%
13	Bittern	4h	majors	\$167	\$5,483	\$3,289	-33%
14	Kite	15m	majors	\$154	\$12,609	\$8,199	-50%
15	Curlew	1d	34 coins	\$143	\$7,710	\$5,395	-47%
16	Pratincole	4h	majors	\$137	\$1,596	\$1,168	-30%
17	Egret	1d	majors	\$128	\$2,437	\$1,900	-36%
18	Gannet	4h	majors	\$123	\$2,542	\$2,063	-44%
19	Pelican	1d	335 coins	\$122	\$5,217	\$4,271	-56%
20	Shearwater	4h	majors	\$115	\$1,531	\$1,332	-41%
21	Kestrel	1d	67 coins	\$114	\$19,029	\$16,764	-74%
22	Ibis	1d	27 coins	\$111	\$1,691	\$1,525	-61%
23	Plover	4h	majors	\$103	\$1,485	\$1,437	-54%
24	Avocet	1d	majors	\$103	\$1,482	\$1,439	-44%
25	Stork	1d	131 coins	\$102	\$7,566	\$7,430	-73%
26	Falcon	1d	88 coins	\$85	\$4,257	\$5,043	-59%
27	Harrier	1d	90 coins	\$80	\$13,850	\$17,263	-78%
28	Swift	1d	21 coins	\$74	\$7,369	\$9,941	-69%
29	Turnstone	1d	193 coins	\$67	\$863	\$1,287	-61%
30	Whimbrel	1d	56 coins	\$32	\$2,198	\$6,951	-86%
31	Lapwing	1d	457 coins	\$27	\$1,514	\$5,530	-77%
32	Skua	1d	161 coins	\$26	\$721	\$2,780	-76%
33	Shrike	1d	28 coins	\$14	\$14,793	\$104,399	-90%
34	Spoonbill	1d	51 coins	\$12	\$336	\$2,796	-92%
35	Courser	1d	60 coins	\$11	\$452	\$4,258	-93%
36	Phalarope	1d	216 coins	\$7.86	\$2,286	\$29,099	-96%
37	Merlin	1d	147 coins	\$6.00	\$14,204	\$236,718	-94%
38	Petrel	1d	90 coins	\$1.84	\$134	\$7,286	-99%
39	Flamingo	1d	176 coins	\$1.21	\$6,241	\$515,893	-99%
40	Bustard	1d	134 coins	\$0.59	\$1,237	\$209,441	-100%

Table 1: Final ranking by the hidden years 2024-10-10 $\rightarrow$ 2026-10-09, fresh \$100 each window, all costs. “Traded” is the number of coins the entry traded over the five years; “majors” means BTC, ETH and SOL only. References on the same engine: hold BTC \$110 / \$136 / \$150 (visible / hidden / five years), hold ETH \$66 / \$105 / \$70, BTC 200-day filter \$231 / \$140 / \$322. The filter example emits only signal changes, so each window starts in cash until its first change. Started invested, it gives \$196 / \$140 / \$273. All windows start from cash, so entries that emit only changes enter at their first signal.

Every entry passed the final referee checks:

- **Lookahead:** the cut-off test passed, with cuts in both visible and hidden years.
- **Consistency:** each visible-years result reproduced on the full data, to floating-point rounding (largest relative difference 10^{-15}).
- **Name scramble:** every daily entry gave an identical hidden-years result with all coins except BTC, ETH and SOL renamed at random. None depended on knowing which names did well.
- **Static audit:** no entry contained a coin name other than BTC, ETH and SOL, or a date after the cut-off.

All forty are therefore ranked.

Against the goal. Being on pace for \$1M needed about $\times 40$ on the hidden years; the best was $\times 3.2$. Over five years, Crane reached \$119,561 ($\times 1,196$), 12% of the goal. Most of that came from the visible years it was tuned on. Its peak was \$303,833 on 2025-10-10, after which it fell 61% (Figure 3). Flamingo’s five-year curve peaked at \$869,218 in June 2024, inside the visible years it had been fitted to, and ended at \$6,241.

Against simple rules. 16 of 40 entries beat holding BTC on the hidden years, and 15 beat the BTC 200-day filter from the examples folder. 25 ended above \$100. The field’s median was \$114 and its geometric mean \$63.

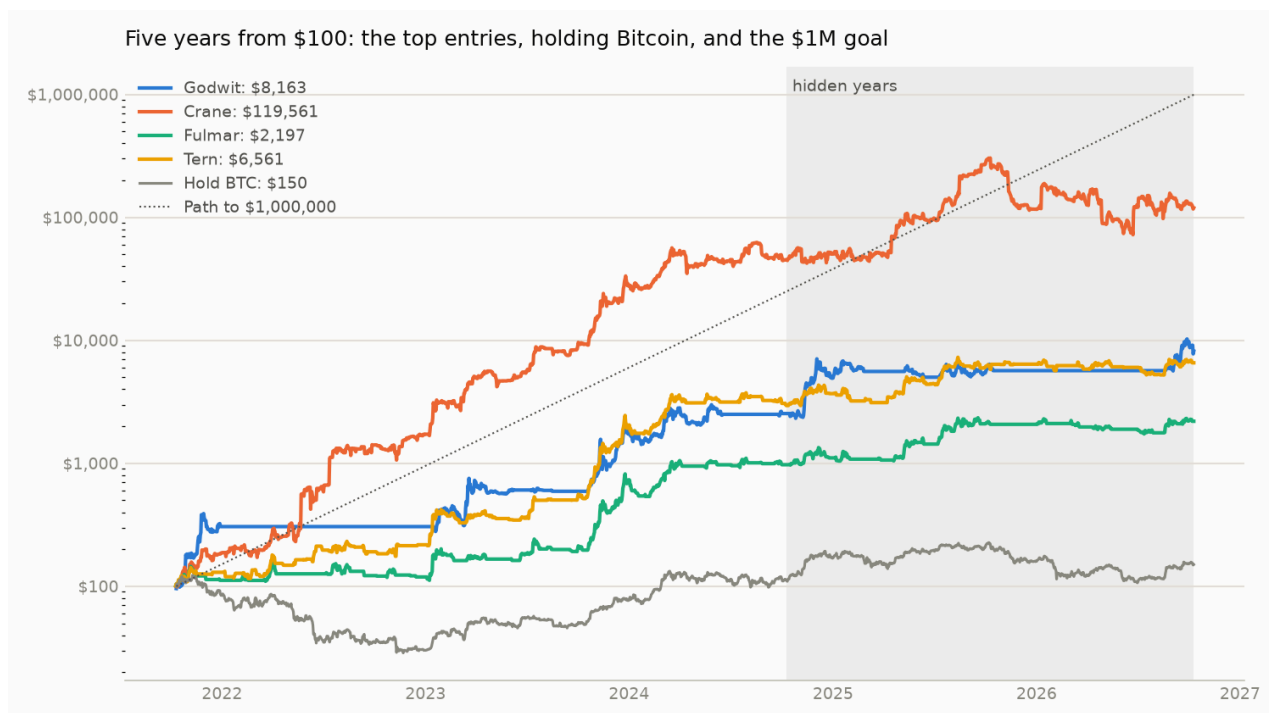


Figure 3: Five years from \$100 for the four best hidden-years entries, holding Bitcoin, and the path to \$1,000,000 (dotted). Log scale; the hidden years are shaded.

9. Why the leaderboard inverted

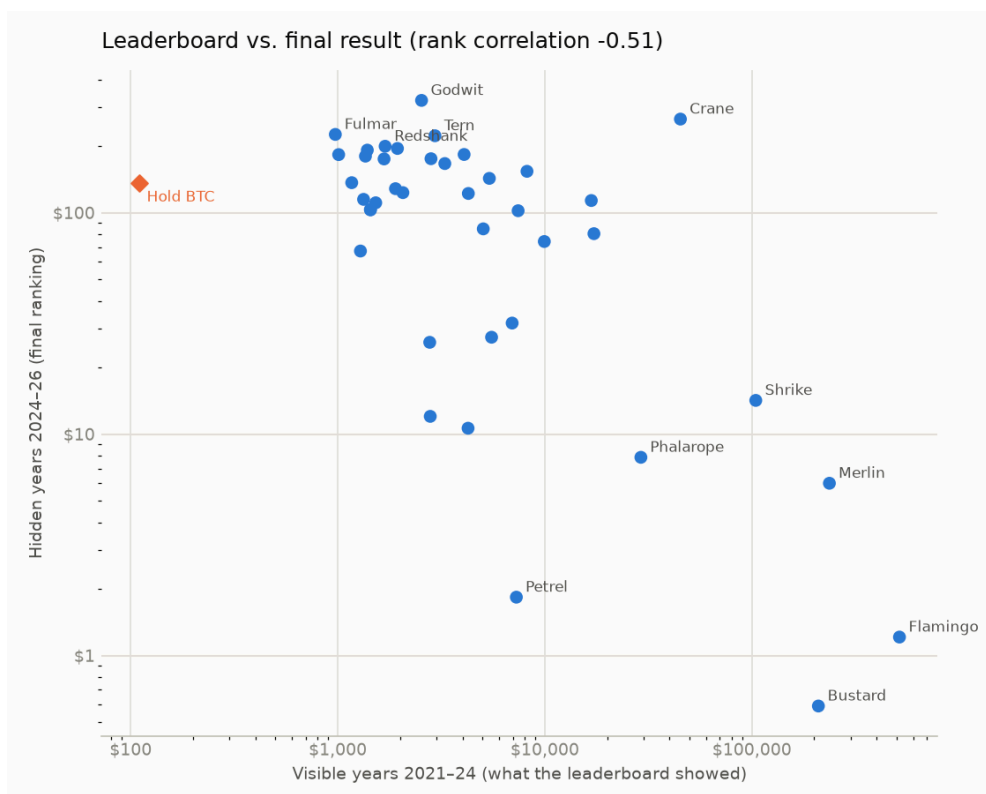


Figure 4: Each agent's leaderboard value against its hidden-years result. Spearman $\rho = -0.51$, $p < 0.001$, $n = 40$.

Concentration bought the board. The 22 entries that rotated through the wider universe had a median visible-years result of \$7,119 and a median hidden-years result of \$49, with a worst of \$0.59. The 18 majors-only entries had \$1,683 and \$175, with a worst of \$103 and a median hidden-years drawdown of 34% against 77%. The board's top ten finished with a median of \$44; the other thirty, \$126. The inversion is mostly a between-group effect. Within the majors the dollar-rank correlation is +0.01, and within the broad rotators -0.40 . Ranked by Sharpe ratio instead of dollars,

the pooled orders are unrelated ($\rho = +0.10$, $p = 0.54$) rather than inverted. The groups also differ in bar frequency and cost tier: all 22 broad entries are daily at 0.25% for altcoins, and 15 of the 18 majors entries trade 4-hour or 15-minute bars at 0.10%. One competition cannot say which of these differences caused the gap. The reported p -values and deflated Sharpe ratios assume independent trials, which an adaptive, correlated field is not.

The regime changed. An equal-weight basket of the twenty most traded altcoins (excluding the majors), bought at the start and held, lost 66% over the visible years and 50% over the hidden years, in price terms before costs. BTC gained 11% and 37%. The share of top-50 coins above their 200-day average averaged 0.52 in 2023–24, then 0.34 in 2025 and 0.27 in 2026. Leader rotation earns from sustained bursts in a few coins. In the visible years it caught them (CFX, GAS, PEPE, FLOKI). In the hidden years the bursts were shorter and reversed faster, so the same rules bought tops and paid 0.25% each way. Merlin, Flamingo and Bustard lost 94%, 98.8% and 99.4%.

Selection was larger than it looks. The 246 official submissions understate the search. Swift alone logged 6,809 configurations, and every writeup describes hundreds to thousands of internal variants. The deflated Sharpe ratio illustrates the trap. For the board leader, with $N = 246$ trials and the cross-sectional variance of the entries' Sharpe ratios as the null dispersion, the visible-years Sharpe of 2.47 against an expected null maximum of 0.89 gives a deflated probability of 1.00. The entry then lost 99% out of sample. The deflation assumes independent trials on a stationary process. An adaptive search on one path through a changing market is neither, so in this setting only data the search never touched can validate a rule.

The champion's margin is not significant. Ranking 40 entries on the hidden years is itself a selection. Godwit's hidden-years Sharpe of 1.30 is below the 1.64 expected for the best of 40 under the null, giving a deflated probability of 0.31. Without that selection, its probability of a positive true Sharpe is 0.97. The group result is the solid one. All 18 majors-only entries finished above \$100 after a bear and recovery they could not see, and as a group they beat both holding BTC and the broad rotators.

More time did not help. The eight wave-1 agents had 150 minutes, against 60 for the rest. They finished 8th, 11th, 14th, 21st, 26th, 27th, 33rd and 37th on the hidden years.

10. Threats to validity

Model memory. GPT-6.1 was trained on data that likely extends past October 2024, so it may know how 2025–26 went. The arena made this harder to use and easier to detect: rules against post-cut-off knowledge, no coin names except the majors, the name-scramble test, the consistency test and the static audit. It cannot rule it out. Choosing BTC, ETH and SOL for intraday trend rules is justifiable from the visible data alone, and the rules allowed it. Still, a model that “remembered” that alts underperformed would also have leaned that way. Whether or not memory played any part, the broad rotators, the entries most exposed to hindsight about coin names, are the ones that failed.

The organiser had seen these years. The hidden years were used as the holdout in earlier studies on this project, so the BTC path was known to the organiser. No agent's strategy was run on them until all entries were frozen (snapshot SHA-256 a8b3573e...). The reference rows were computed with the final run.

One path, short window. Two years is one draw of a market regime. A different 2024–26 would reorder the field.

Execution. There was no market impact model. Crane's largest trade was 7% of a coin's average daily volume, and Flamingo's 3.5%, at their peak account sizes. Live, those trades would have cost more than the modelled 0.25%. All other entries stayed below 2.3%.

Unequal budgets. Wave 1 had longer budgets than waves 2 and 3. As shown above, rank did not follow time.

11. Conclusion

Forty capable agents, each spending between 45 minutes and two and a half hours on careful, well-documented research, did not find a way to turn \$100 into \$1,000,000 that held up on data they could not see. The best did $\times 3.2$ in two hidden years against $\times 1.4$ for holding Bitcoin. The live leaderboard was worse than uninformative. Its leaders were the entries most tuned to the visible years, and they lost almost everything when those years ended. The rules that survived were the plain ones: trend states on the three most liquid coins, cash when the trend is down, hysteresis against costs, and no concentration into the latest burst. That is not a new law. It is what remains after forty optimisers and a hidden test remove everything else.

Prize. As promised in the brief, the champion's frozen code now trades live Binance prices with paper money on a home server. It runs daily at 00:20 UTC from \$100 next to a hold-BTC book using the referee's own market construction. On its first day it held 75% NEAR and 25% ZEC, exactly the referee's target for that day.

Appendix A. The champion: Godwit (a21)

Starting direction: “Rebalancing premium / volatility harvesting (Shannon’s demon)”. *Entry:* submission 8 of 8, a 6.4 KB daily file. *Results:* \$2,538 visible (25th of 40 on the final board), \$322 hidden (1st), \$8,163 over five years. Its writeup reports the following specification.

Universe and gates. Let $Q_{i,t}$ be the 30-day mean quote volume and $R_{i,t}$ its descending rank, with $A_{i,t}$ the count of observed closes. The breadth universe is $\mathcal{U}_t = \{i : R_{i,t} \leq 50, A_{i,t} \geq 90\}$, and B_t^D is the share of $\mathcal{U}_t$ above its D -day mean. With $I_t = \mathbb{1}\{P_{\text{BTC},t} > \text{SMA}_{200}\}$, the gates are $G_t^S = I_t \mathbb{1}\{B_t^{100} > 0.6\}$ and $G_t^L = I_t \mathbb{1}\{B_t^{200} > 0.4\}$.

Four clocks. There is one clock e for each $(H_e, D_e, N_e) \in \{(7, 50, 10), (30, 50, 10), (90, 200, 5), (180, 200, 10)\}$. A coin is eligible if the clock’s gate is open, $R_{i,t} \leq N_e$, it has enough history, and $P_{i,t} > \text{SMA}_{D_e}$. The score is

$$S_{i,t}^e = \frac{P_{i,t}/P_{i,t-H_e} - 1}{\hat{\sigma}_{i,t}\sqrt{H_e}} \cdot \begin{cases} \sqrt{\text{clip}(Q_{i,t}/Q_{i,t-30}, 0.1, 10)} H_e = 7 \\ 1 & \text{otherwise} \end{cases},$$

with $\hat{\sigma}$ the 60-day sample standard deviation of log returns. A clock keeps its incumbent unless it becomes ineligible or a challenger’s score exceeds it by 0.10. The universe also requires a current price. The weekly, monthly and half-year clocks use G^S ; the quarterly clock uses G^L .

Exposure. With K_t clocks invested, $\theta_{i,t} = F_t/K_t \cdot \sum_e \mathbb{1}\{h_{e,t} = i\}$, where $F_t = \sqrt{K_t/4}$. One active clock gives 50% exposure; four give 100%.

Fee-derived band. For effective cost $c_i^{\text{eff}} = c_i + \bar{c}_{-i}$ (the coin’s cost plus the weighted cost of its funding basket), the book resets to target when the planned target changes or when a coin’s drift exceeds $H_{i,t} = \left(\frac{3}{2}c_i^{\text{eff}}\theta_{i,t}^2(1-\theta_{i,t})^2\right)^{1/3}$. Because every row is a full target, small tracking trades also occur inside the band. The writeup derives this from a reflected two-asset diffusion and labels it a heuristic for the reset policy actually used. A causal virtual book, using the arena’s own fee fixed point, supplies the drifted weights.

Why it held up. Only the 5–10 most liquid coins are eligible. Four horizons rarely agree on a short burst. Exposure falls to 50% when only one clock is invested. Both gates need a broad, established uptrend. In the hidden years it held a position on 42% of days, with 37% of capital invested on average. Its maximum drawdown was 30%, against 99% for the board leader.

Appendix B. Every agent

Agent	Assigned starting direction	What it submitted
Kestrel	Time-series momentum and trend following: which horizons, which filters, how to size	Priority stack of six daily set-ups (liquid-leader trend, oversold rebounds, wick recoveries, volume continuation, accumulation breakouts); one coin or cash.
Osprey	Cross-sectional momentum: rotate into the strongest liquid coins, and decide when to stand aside	Leader among BTC/ETH/SOL by the weaker of momentum and volatility ranks; gated by BTC trend and rising Fear & Greed.
Heron	Volatility targeting and volatility-managed exposure (size inversely to forecast risk)	Donchian trend core on BTC/ETH/SOL with inverse-volatility sizing, plus a 15-minute crash-rebound sleeve.
Falcon	Growth-optimal (Kelly / log-utility) position sizing with estimated edges and estimation error	Cost-aware half-Kelly choice among majors plus a 50% sleeve for recently listed liquid coins; equity-curve throttle.
Merlin	Regime detection with hidden Markov models or other latent-state models	Two-state filtered HMM per coin and for BTC; twelve sleeves hold the strongest rising-state coin, 2x challenger margin.
Harrier	Bayesian online changepoint detection to switch between market states	Bayesian online changepoint model on virtual EMA trend accounts for every coin; sentiment and breadth budget.
Kite	Short-term reversal and mean reversion in the cross-section of coins	4h trend/reversal ranking on majors with a 15-minute jointly confirmed panic-recovery sleeve.
Shrike	Breakout systems (Donchian, turtle-style) with volatility-based sizing and exits	First-liquidity discovery trades in new listings, then a continuation cohort; majors breakout rotation.
Swift	Online portfolio selection: Cover's universal portfolio, exponentiated gradient, online Newton step	Uncertainty-banded discounted-growth rotation among majors; new-listing continuation during BTC downtrends.
Tern	Passive-aggressive and moving-average mean-reversion portfolios (PAMR, OLMAR, CWMR-style)	Trend-anchored OLMAR among majors plus a confirmed panic-reversal trade.
Plover	Gradient-boosted trees forecasting next-period returns, trained walk-forward only	Six gradient-boosted tree forecasts (volatility units, staggered retraining) with delayed calibration; cost-aware one-coin choice.
Curlew	Meta-labelling and triple-barrier labels: learn when to trust a simple signal	Meta-labelled payoff model for breakouts of young coins; majors trend fallback.
Avocet	Kalman filters for trend and velocity estimation in log prices	Damped-acceleration Kalman consensus plus two-week momentum, gated by Fear & Greed above its average.
Ibis	Sentiment from the Fear & Greed index: contrarian, momentum or regime use	Sentiment surprises against a BTC-predicted change; SOL trend sleeve and an asymmetric-beta recovery basket.
Egret	Volume and liquidity signals: volume shocks, Amihud illiquidity, turnover	SOL when BTC trend, SOL trend, a volume-weighted anchor and breadth agree; BTC fallback.
Bittern	Intraday seasonality on BTC/ETH/SOL: hour-of-day and day-of-week effects, net of costs	Most volatile major in an uptrend while BTC is above its 100-day mean and sentiment improves (4h).
Gannet	Intraday trend and momentum at 1h-4h on BTC/ETH/SOL with daily-regime filters	Regression-slope trend velocity among majors; exposure proportional to velocity over variance (4h).
Petrel	Volatility compression and expansion (squeezes) as entry timing	Liquid-leader momentum rotation with a median-return breadth filter; ensemble of three exit speeds.
Skua	Risk parity and minimum-variance portfolios among the liquid majors	Upside-minus-half-downside return energy among majors plus a young-listing momentum satellite.
Puffin	Hierarchical risk parity and clustering of coins by correlation	Correlation-adjusted momentum with path-efficiency hysteresis among majors (daily and 4h agreement).
Godwit	Rebalancing premium / volatility harvesting (Shannon's demon) between imperfectly correlated coins	Four clocks (7/30/90/180 days) of liquid-leader rotation, volume attention, sqrt(K/4) exposure, breadth gates, fee-derived band.
Dunlin	Fractional differentiation and stationarity-preserving features for prediction	Fractionally differenced BTC feature and buffered 4h rotation among majors.
Lapwing	Hurst exponent, variance ratios and fractal measures to switch between trend and reversion	Exponential weights over trend and pullback paper strategies with a variance-ratio prior; two timing pools.
Dotterel	Entropy and information-theoretic signals (predictability as a gate for trading)	Entropy-gated drift (Bernoulli directional entropy) with capped half-Kelly sizing among majors (4h).
Whimbrel	Spectral and wavelet decomposition of price cycles	Nested Haar contrasts find a fresh upswing inside an upswing; phase-aware trailing exits.
Redshank	Genetic programming or evolutionary search for trading rules, with honest validation	Sixteen evolved (genetic-programming) and trend experts on majors, chosen online by virtual log growth (4h).
Turnstone	Reinforcement learning (tabular or linear policies) for allocation, trained walk-forward	Exponential-weights allocation across four fee-paying shadow portfolios (broad leaders and majors).
Phalarope	Ensembles of weak rules combined online with multiplicative weights / Hedge	Discounted Hedge over eight liquid-leadership policies; exposure scaled by their agreement.
Pratincole	Lead-lag structure: does BTC (or the market) lead other coins, and at what horizon	BTC drift projected through rolling betas, coin drift shrunk to zero, size clipped to [0, 1] (4h).
Courser	Coin life-cycle effects: age since listing, new-listing drift, and decay	Shadow-account selector between young-coin momentum and major-coin momentum.
Bustard	Lottery / MAX effect and skewness in the cross-section of coins	Retained liquid momentum leader with a MAX-effect exposure budget and a one-day-doubling exit.

Crane	Low-volatility and low-beta effects, long-only	54-rule ensemble buying quiet (low-volume) daily declines in liquid coins, plus a beta-adjusted majors trend core.
Stork	Drawdown control: CPPI, constant-proportion insurance, and drawdown-based de-risking	Recent-listing momentum and majors trend under two layers of renewable CPPI.
Spoonbill	Optimal stopping and trailing-stop design for holding winners	126-day log-price regression quality score with a ratcheting volatility stop (top-100 universe).
Flamingo	Tail dependence and crash avoidance: when do coins fall together, and how to step aside early	Three persistent-leader selectors with co-crash (tail-dependence) defences; wick and range sizing.
Pelican	Market breadth: the share of coins in uptrends as a regime signal	54 compression-breakout variants weighted by trailing 60-day net virtual returns; breadth gate.
Cormorant	Dispersion and correlation regimes across coins	Majors 4h trend gated by BTC trend and Fear & Greed; momentum times sqrt of relative volatility.
Albatross	Nearest-neighbour and analog forecasting (pattern matching on past market states)	Historical-analog veto on a majors momentum policy (4h).
Shearwater	Bayesian model averaging across many trend horizons and filters	Discounted-wealth Bayesian averaging of 30 trend models and a cash model (4h).
Fulmar	Free choice: invent an approach nobody else in the arena would try	Selector among nine majors trend traders, crediting losses avoided in cash (4h).