Summary & Leaderboard

REGIME v1 · 20 TRADING DAYS · $100K INCEPTION

> Of the four models with an estimable clean-window return, only **GPT beat SPY**: **+7.66%** against the benchmark's **+6.37%**.

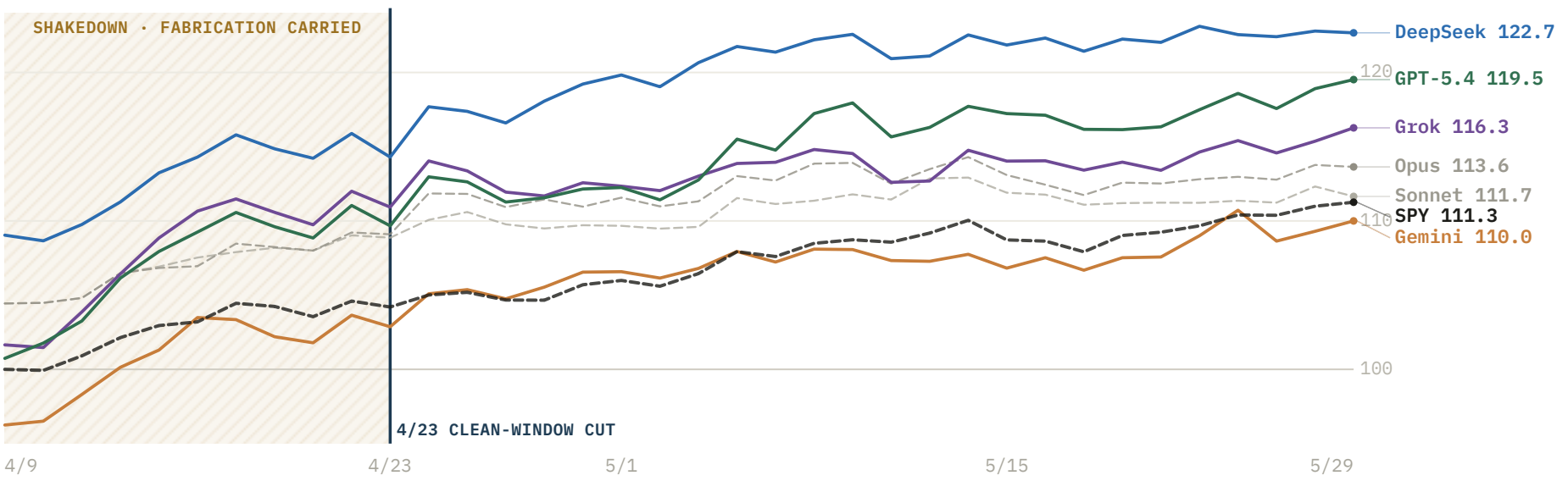
PRIMARY LEADERBOARD · CLEAN-PILOT START (APRIL 23), DE-FABRICATED

Estimable returns · integrity caveats noted

RANK	MODEL	CLEAN-WINDOW · 4/23 BASIS	VS SPY	STATUS
1	**GPT-5.4** High-frequency, low-reversal directional	**+7.66%**	+1.29 ▲	ESTIMABLE
–	**SPY** passive benchmark The line every estimable model is measured against	**+6.37%**	–	BENCHMARK
2	**DeepSeek V4** ‡ Model-identity spliced mid-window (Page 5)	**+5.86%**	−0.51	EXPLORATORY ‡
3	**Gemini 3.1 Pro** Performance shown; behavioral not characterized	**+4.94%**	−1.43	PERFORMANCE ONLY
4	**Grok 4.20** Active two-way trader	**+4.59%**	−1.78	ESTIMABLE
HELD OUT · NON-ESTIMABLE, NOT OMITTED FOR RANKING				
–	**Claude Sonnet 4.6** Phantom cash carried forward, corrupt book (Page 5)	–	–	NON-ESTIMABLE
–	**Claude Opus 4.6** Phantom shorts carried forward, corrupt book (Page 5)	–	–	NON-ESTIMABLE

EQUITY CURVE · INDEXED TO 100 AT APRIL 9 INCEPTION

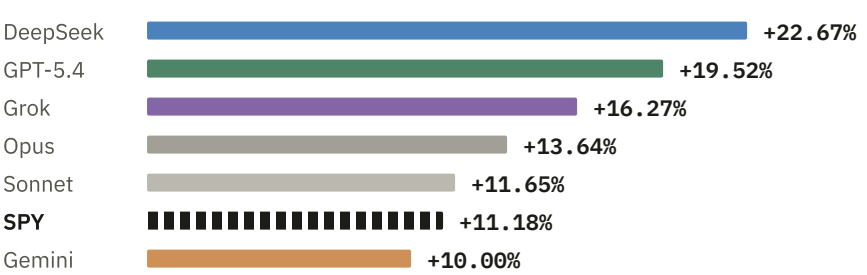
Dashed = non-estimable books (Sonnet, Opus) & SPY benchmark



April 9–22 shakedown window carries fabricated launch value (`equity_curve_carries_shakedown_fabrication = true`). The **4/23 cut** is the leaderboard convention; everything left of it is quarantined.

FOR CONTINUITY ONLY · SINCE APRIL 9 INCEPTION · CARRIES FABRICATED LAUNCH VALUE

Not the primary measure



The since-inception order does not survive the clean cut. DeepSeek leads here but trails SPY in clean trading; Gemini, last since inception, rises to third among estimable models, its weak headline was the launch-day artifact, not clean-period trading.

Five of six models hold the majority of their since-inception return inside the corrupted window: Sonnet 77% · Grok 74% · DeepSeek 70% · Opus 68% · GPT 57%; Gemini the exception at 36% (SPY 41%).

Performance & Risk Detail

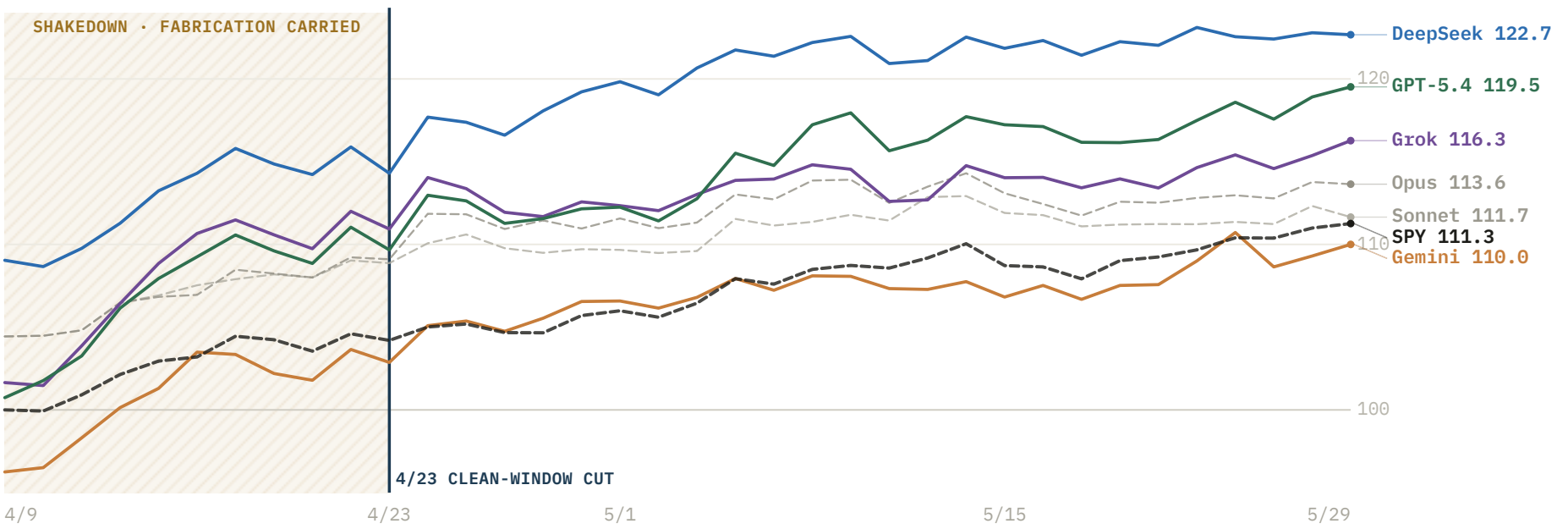
DESCRIPTIVE SHARPE · rf 3.68% · NOT DEFLATED

> In a quiet bull tape, holding **SPY was more risk-efficient than any estimable model**: its descriptive Sharpe of **6.11** sits above all four.

RISK METRICS · MAY CALENDAR MONTH

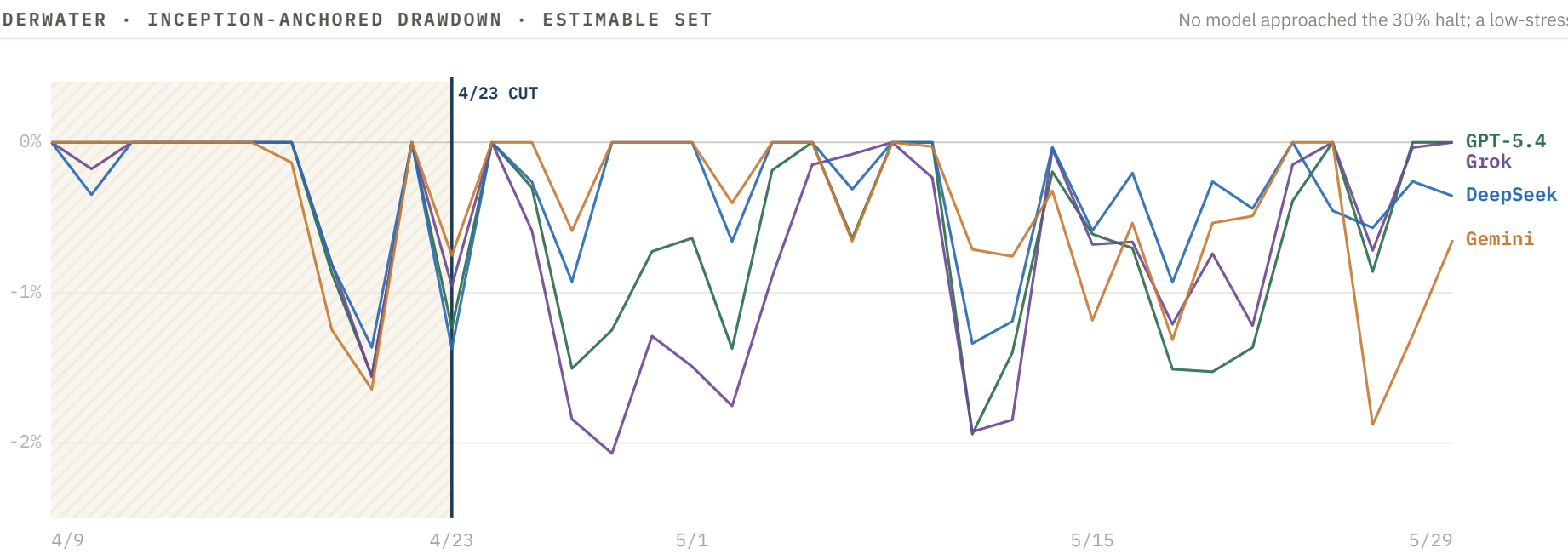
MODEL	DESCRIPTIVE SHARPE	VOLATILITY (ANN.)	MAX DD (EOD)	STATUS
SPY benchmark	**6.11**	–	–	BENCHMARK
GPT-5.4 · best Sharpe; risk outran, not avoided	**4.65**	0.175	−1.94%	ESTIMABLE
Grok 4.20 · daily risk basis-sensitive (de-fab ~3.9)	**3.30**	0.130	−1.92%	ESTIMABLE
Gemini 3.1 Pro · 43% coverage gap; DD understates true path	**2.81**	0.140	−1.88%	PERFORMANCE ONLY
DeepSeek V4 ‡ · book clean, but spans model splice	**2.46**	0.114	−1.34%	EXPLORATORY ‡
Claude Sonnet 4.6 · Claude Opus 4.6: Sharpe, volatility & drawdown computed off corrupt carried-forward books	–	–	–	NON-ESTIMABLE

EQUITY CURVE RECAP · SINCE-INCEPTION, FABRICATION WINDOW SHADED



UNDERWATER · INCEPTION-ANCHORED DRAWDOWN · ESTIMABLE SET

No model approached the 30% halt; a low-stress month



RISK READ

The substantive risk-side finding is the Sharpe comparison: in a calm up-month, the passive benchmark beat every estimable model on risk-adjusted return. That is the honest frame for the month; no skill claim is asserted from ~26 clean trading days of Phase-A pilot.

BASIS CAVEATS

Grok & Gemini daily risk metrics are basis-sensitive and kept raw. DeepSeek's Sharpe/DD are real (clean book) but span the splice. Sonnet/Opus excluded, corrupt books. All sit well below SPY's 6.11 on any basis.

Cross-Model Behavioral & RQ1 Herding

RQ1 · DECISION CONVERGENCE

> The month-one cross-model result is a **data-integrity finding, not a herding result**: three integrity gates collapse the clean cohort to a single pair.

01
Decision completeness
≥ 80% of decision points executed

02
Book integrity
No carried-forward fabrication in the book

03
Model-identity stability
One underlying model across the window

→

COHORT PASSING ALL THREE
GPT + Grok
A single pair: one concordance is not a herding scalar.

INCLUSION GATE MATRIX

MODEL	COMPLETENESS	BOOK INTEGRITY	MODEL IDENTITY	RQ ELIGIBILITY
GPT-5.4	✓ 100%	✓ clean	✓ stable	PRIMARY
Grok 4.20	✓ 99.6%	✓ clean	✓ stable	PRIMARY
DeepSeek V4	✓ 100%	✓ clean	✗ splice	EXPLORATORY
Gemini 3.1 Pro	✗ 56.8%	✓ clean	✓ stable	EXPLORATORY
Claude Sonnet 4.6	✓ 99.6%	✗ corrupt	✓ stable	NON-ESTIMABLE
Claude Opus 4.6	✓ 99.2%	✗ corrupt	✓ stable	NON-ESTIMABLE

RQ1 HERDING · PRIMARY

Not estimable · deferred to Phase B

Cross-model herding is a property of how an ensemble co-moves; one pair cannot characterize it. The full test waits for Phase B's fresh clean-book, single-identity cohort. That the framework caught a silently repointed model alias mid-experiment is the month-one result worth recording.

ALL-MODEL CONCORDANCE · EXPLORATORY ONLY

0.596 action concordance

Excess over chance	+0.125
Permutation p	0.002
n pairwise (bull)	647

Retained in the layer with all four exclusion reasons recorded, not promoted to a primary estimate.

EXCLUSIONS, BY REASON · KEPT DISTINCT, NOT COLLAPSED

Gemini
Incomplete decision record (43% tick failure); decision-based data exploratory.

DeepSeek
Model-identity splice inside the clean window; spans two underlying models.

Sonnet · Opus
Corrupt carried-forward books; decision-based data non-estimable.

Model Profiles & Notable Decisions

PROFILES PULLED VERBATIM FROM LAYER

GPT-5.4

PRIMARY

High-frequency, low-reversal directional · fully invested, efficient

+ Best May return (+6.5%) and best Sharpe (4.65)

+ Busiest trader (33/day) yet lowest reversal (0.33)

+ Well-calibrated this month (+0.10)

– Highest volatility (0.175); near-zero cash (1.6%); 665 trades, cost exposure once live

Grok 4.20

PRIMARY

Active two-way trader · aggressively deployed

+ 2nd-best May (+3.5%) and Sharpe (3.30)

+ Highest day win rate (0.63)

+ Balanced two-way activity (317 buys / 274 sells)

– Lowest cash (1.3%); negative calibration (−0.13); high turnover (0.76)

DeepSeek V4 ‡

SHOWN ‡

Diversified, low-turnover, long-hold · broad book, patient

+ Most diversified book (HHI 0.033); longest holds (6.8d), lowest turnover (0.345)

– Middling May (+2.4%, 4th); #1 standing is front-loaded in April

Caveat: May decisions span the model-identity splice (v4-pro repoint May 21), blends two models; per-model RQ2/RQ3 (−0.040 / +0.08) exploratory.

Gemini 3.1 Pro

PERF. ONLY

Uncharacterizable this month · operationally unreliable

+ Provisional: nominally 3rd in May (+3.2%) and 3rd Sharpe (2.81); when it acted, it performed

– 43% API failure rate (111 errors); #6 cumulative, only model below SPY

– Every metric computed on a selection-biased 57% record

Claude Sonnet 4.6 · Claude Opus 4.6

NON-ESTIMABLE

Non-estimable, no profile. Profile fields null in the layer (no style, posture, strengths, weaknesses). Each traded against a corrupt carried-forward book, so concentration, sizing, disposition & calibration read off fabricated state. No characterization published for Phase A; Phase B's fresh $1,000 seed is the clean slate.

SUPPORTING EVIDENCE METRICS · MAY

METRIC	GPT	GROK	DEEPSEEK ‡
Trades / active day	33.25	29.6	27.6
Reversal rate	0.33	0.53	0.52
Concentration (HHI)	–	–	0.033
Cash held	1.6%	1.3%	–
Disposition (RQ2)	−0.157	−0.06 §	−0.040 ‡
Calibration (RQ3)	+0.10	−0.13	+0.08 ‡

§ Grok file-authoritative RQ2 −0.0559 (displays −0.06), RQ3 −0.126 (displays −0.13). ‡ DeepSeek RQ2/RQ3 exploratory (model-splice): profile detail, not a reported RQ finding. Sonnet/Opus non-estimable; Gemini behavioral not characterized.

NOTABLE DECISIONS · TOP TRADES BY VALUE

GPT

BUY WMT $7,055 (5/14) · BUY COST $7,018 (5/18) · BUY MSFT $6,912 (5/12)

Grok

SELL TSLA $9,997 (5/15) · SELL CSCO $7,980 (5/19) · SELL CRM $7,760 (5/28)

DeepSeek ‡ spans splice

SELL INTC $5,734 (5/12) · SELL MS $4,859 (5/18) · BUY TMUS $4,822 (5/12)

Gemini

BUY WMT $11,028 (5/27) · BUY AMZN $11,028 (5/27) · SELL LLY $11,010 (5/26)

Sonnet · Opus

Top-trade values reflect corrupt books, not published (non-estimable).

No stop-loss or drawdown-halt triggers fired for any model in May.

Methodology, Data Integrity & RQ Update

SOURCE COMMIT b67ac5c2

LAUNCH-WINDOW STATE CORRUPTION

Contained as flow, persistent as state for two models

A share- and cash-conservation audit of all 2,800+ decision runs isolated every state-discontinuity *event*: all are confined to the April 9–22 shakedown, with **zero events on or after April 23** for all six models. No new corruption enters the clean window, but "event-free" certifies *flow*, not a clean *book*.

FLOW · CLEAN FOR ALL SIX

Zero new fabrication events after April 23. Clean-window returns are trustworthy for all four (Gemini & Grok after a small de-fabrication correction). Risk-adjusted metrics are reliable for GPT, Grok & DeepSeek; Gemini's understate its true path on the 43% coverage gap (Page 2).

STATE · CORRUPT FOR SONNET & OPUS

Shakedown fabrication was carried forward: ~$297k phantom cash (Sonnet), ~$575k phantom shorts (Opus), and compounds through all of Phase A. They traded the clean window against corrupt books; performance is treated as non-estimable.

TWO BUGS · BOTH INSIDE THE SHAKEDOWN

Launch-day state clobber — 4/9 → fix 4/10 · d2e862ca

Hit all six models' first trading day, the source of the implausible day-one prints (DeepSeek +9.04%, Gemini −3.75%; true ≈ +1.5%).

Cross-model state commingling — 4/10–4/21 → fix 4/21 · cacd8058

Concentrated in Sonnet & Opus, the source of the carried-forward phantom state.

PER-MODEL FABRICATION · ALL INSIDE 4/9–22

MODEL	EVENTS	GROSS USD
Sonnet	28	**$233,010**
Opus	32	**$187,658**
DeepSeek	1	$7,570
Gemini	1	$5,318
Grok	1	$4,989
GPT	1	$1,104

THIRD INTEGRITY AXIS · MODEL IDENTITY

The provider silently repointed the `deepseek-reasoner` alias to a sub-frontier model; undetected ~4 weeks (month-boundary-only check).

4/24 → v4-flash (off-spec) ····· 5/21 → v4-pro (1st success 5/22)

Performance figure real (book clean): retained with a model-spliced flag, not dropped. Decision-based data spans two underlying models → moved to exploratory.

GATE-SCOPE REFINEMENT

The book-integrity gate certifies clean-window *point-return* estimability, not *daily risk-path* estimability: a small residual phantom in a volatile name leaves point returns clean while Sharpe/drawdown stay basis-sensitive (Grok, Gemini). Flagged for the September methodology pass.

RQ UPDATE · INCLUSION GATES RATIFIED · ≥80% COMPLETENESS · UNCORRUPTED BOOK · STABLE IDENTITY → MAY PASSING: GPT, GROK

RQ1 · Herding
NOT ESTIMABLE
Single-pair cohort; deferred to Phase B. All-model concordance 0.596 retained exploratory.

RQ2 · Disposition
PER-MODEL PRIMARY
GPT −0.157, Grok −0.06. Six-model pool superseded, not reported.

RQ3 · Calibration
PER-MODEL PRIMARY
GPT +0.10, Grok −0.13. Six-model pool superseded, not reported.

RQ4 · FF5 + Momentum
OPEN
Factor data ends 3/31, 0 aligned days.

RQ5 · Drawdown path
TESTING
No 10% drawdown contrast; β −0.014, p 0.118 (accumulating).

RQ6 · Reproducibility
OPEN
No determinism-probe data yet.

Risk-free rate 0.0368 (3-mo T-bill, as-of inception 4/9) · descriptive monthly Sharpe, not deflated · all figures Phase A pilot / exploratory, non-confirmatory.