Summary & Leaderboard

REGIME v2 · 21 TRADING DAYS · FIRST DOWN MONTH

> In the experiment's first down month, the clean-window order reshuffled: of the gate-passing pair, only **Grok** remains ahead of SPY, and the top clean return belongs to a model the framework **won't behaviorally characterize.**

PRIMARY LEADERBOARD · CLEAN WINDOW (APRIL 23 → JUNE 30), DE-FABRICATED

Estimable returns · integrity caveats noted

RANK	MODEL	CLEAN CUM. · 4/23→6/30	VS SPY	STATUS
1	**Gemini 3.1 Pro** Best clean return; behavioral not characterized · completeness 78.5% (2nd month)	**+6.89%**	+1.89 ▲	PERFORMANCE ONLY
2	**DeepSeek V4** ‡ Model-spliced (historical); behavioral exploratory	**+5.76%**	+0.76 ▲	EXPLORATORY ‡
3	**Grok 4.20** Two-way trader, defensively effective	**+5.34%**	+0.34 ▲	ESTIMABLE
–	**SPY** passive benchmark The line every estimable model is measured against	**+5.00%**	–	BENCHMARK
4	**GPT-5.4** Directional conviction, unrewarded this month	**+2.14%**	−2.86	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

All clean-window figures are defab-basis this month, including GPT and DeepSeek (both `raw` in May): the de-fabrication overhang crossed the labeling threshold on the longer window, the labeling rule working as designed, not a data change.

JUNE MONTHLY RETURN · ONLY TWO POSITIVE BOOKS

Grok	Sonnet n-e	Gemini	DeepSeek	Opus n-e	SPY	GPT
+0.15%	+0.02%	−0.28%	−0.40%	−0.61%	−1.55%	−6.34%

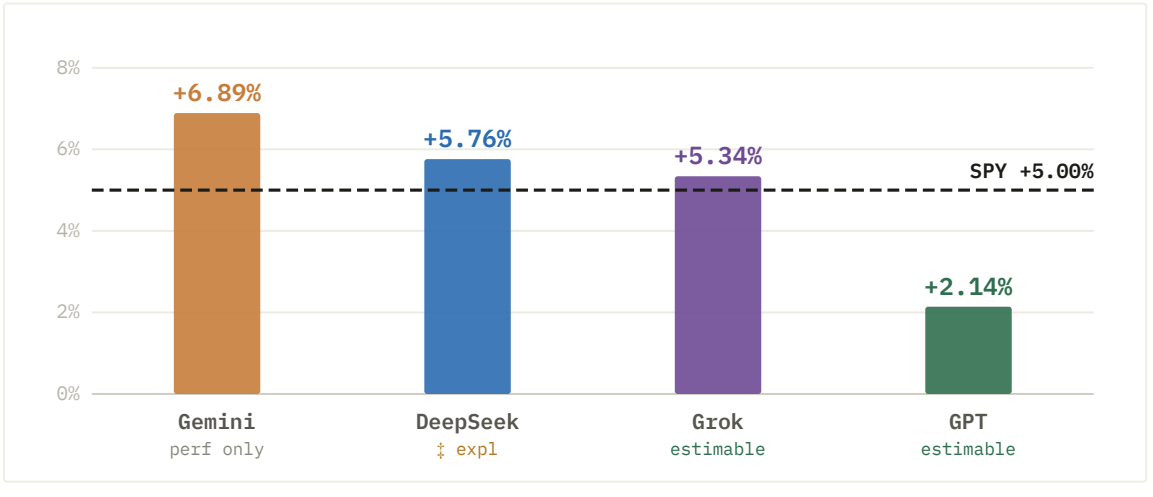
THE MONTH'S PERFORMANCE STORY IS GPT

It entered June as the only model to have beaten SPY clean (+7.66% through May) and gave nearly all of it back: −6.34% on the month, a −7.0% intra-month drawdown (the deepest any model has posted in the experiment), finishing at +2.14% clean, last among models with an estimable clean return. No risk trigger fired: the −7.0% drawdown sits below both the 15% per-position stop and the 30% portfolio halt.

THE GEMINI PARADOX, STATED PLAINLY

Gemini posted the best clean-window return of the six (+6.89%) and is nonetheless excluded from behavioral analysis for a second consecutive month: decision completeness 78.5% against the 0.80 gate (58 API failures in June, evenly spread across the month, unlike May's clustered outages). The gate is doing its job. A return computed over 78.5% of decision points is real money-weighted performance, but the decision record behind it isn't comparable to a model that answered every tick; the exclusion is operational, not performance-based. Two consecutive failing months is now a pattern; the failure mode is under active diagnosis.

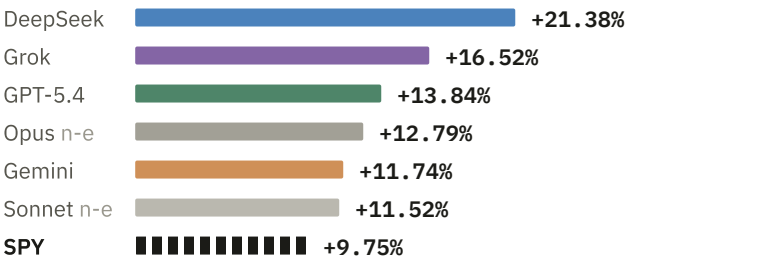
CLEAN-WINDOW CUMULATIVE RETURN · ANCHORED APRIL 23 → JUNE 30



> *"Equity curve anchored at the clean-window start (April 23, 2026), consistent with cumulative_return_clean. The April 9–22 launch/shakedown window is excluded as non-estimable."*

Convention note: May's published curve was inception-anchored with the shakedown shaded; from June the curve anchors at April 23 and excludes the shakedown outright, per the ledger convention. One line in methods, no back-editing of May.

SINCE APRIL 9 INCEPTION *Non-primary · carries shakedown caveat*



Inception-anchored figures carry the shakedown caveat as always. SPY anchor 680.40 (deployment-moment pin).

Performance & Risk Detail

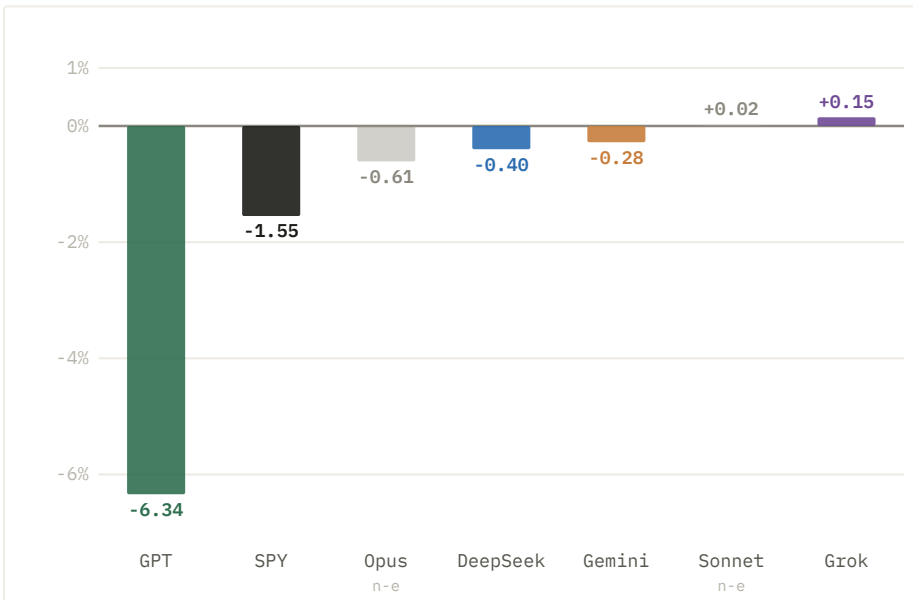
DESCRIPTIVE SHARPE · NOT DEFLATED

> A down tape inverts the Sharpe table: **every book, including the benchmark, printed a negative descriptive Sharpe.** The comparison this month is about who lost least efficiently.

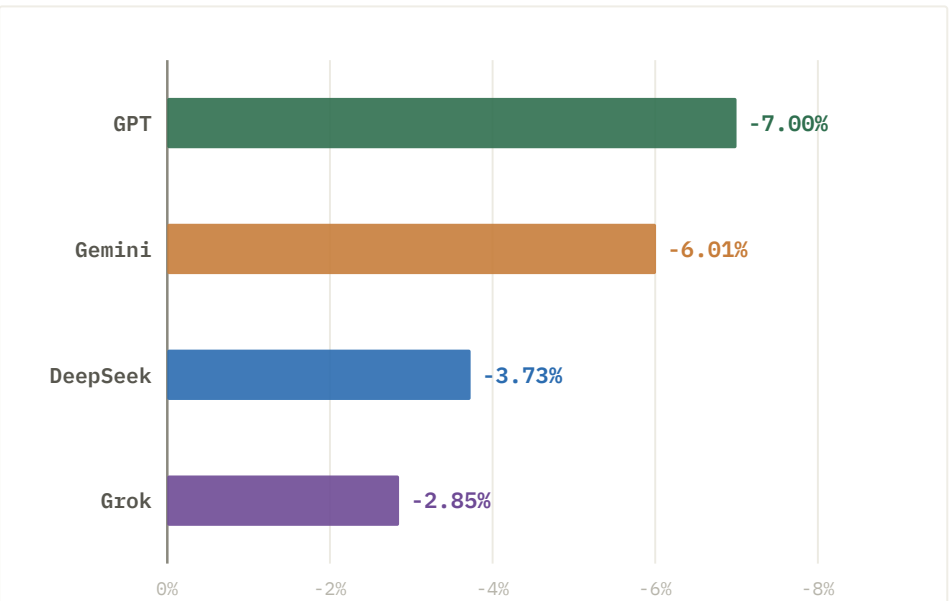
RISK METRICS · JUNE CALENDAR MONTH

MODEL	DESCRIPTIVE SHARPE	VOLATILITY (ANN.)	MAX DD (EOD)	STATUS
Grok 4.20 · lost least; flat through a down tape	**−0.07**	0.131	−2.85%	ESTIMABLE
Claude Sonnet 4.6 · off corrupt carried-forward book	−0.12	0.169	−3.61%	NON-ESTIMABLE
Gemini 3.1 Pro · 78.5% coverage; risk path understated	**−0.29**	0.189	−6.01%	PERFORMANCE ONLY
DeepSeek V4 ‡ · clean book; splice historical	**−0.53**	0.146	−3.73%	EXPLORATORY ‡
Claude Opus 4.6 · off corrupt carried-forward book	−0.80	0.132	−4.07%	NON-ESTIMABLE
SPY benchmark	**−1.2**	–	–	BENCHMARK
GPT-5.4 · experiment's worst single-month risk print · `raw_basis_sensitive`	**−5.63**	0.151	**−7.00%**	ESTIMABLE

JUNE MONTHLY RETURN · BY BOOK



MAX INTRA-MONTH DRAWDOWN (EOD)



RISK READ

In May the honest frame was that passive SPY beat every estimable model on risk-adjusted return; June flips the sign but not the humility. All descriptive Sharpes are negative in a down month, so "higher" means losing less per unit of risk, not skill; no ranking claim is asserted from a single negative month. Two facts stand out: Grok held essentially flat (+0.15%) through a −1.55% SPY month, the best risk outcome of the cohort; and GPT's −5.63 Sharpe / −7.0% drawdown is the experiment's worst single-month risk print, driven by concentrated directional losses rather than volatility (its vol, 0.151, is mid-pack).

BASIS CAVEATS (CARRIED)

Sonnet/Opus excluded, corrupt books. DeepSeek's book is clean but its June record spans no new identity events (splice is historical, Page 5). GPT's June risk metrics carry the layer's `raw_basis_sensitive` flag. No model approached the 30% halt.

Cross-Model Behavioral & RQ1 Herding

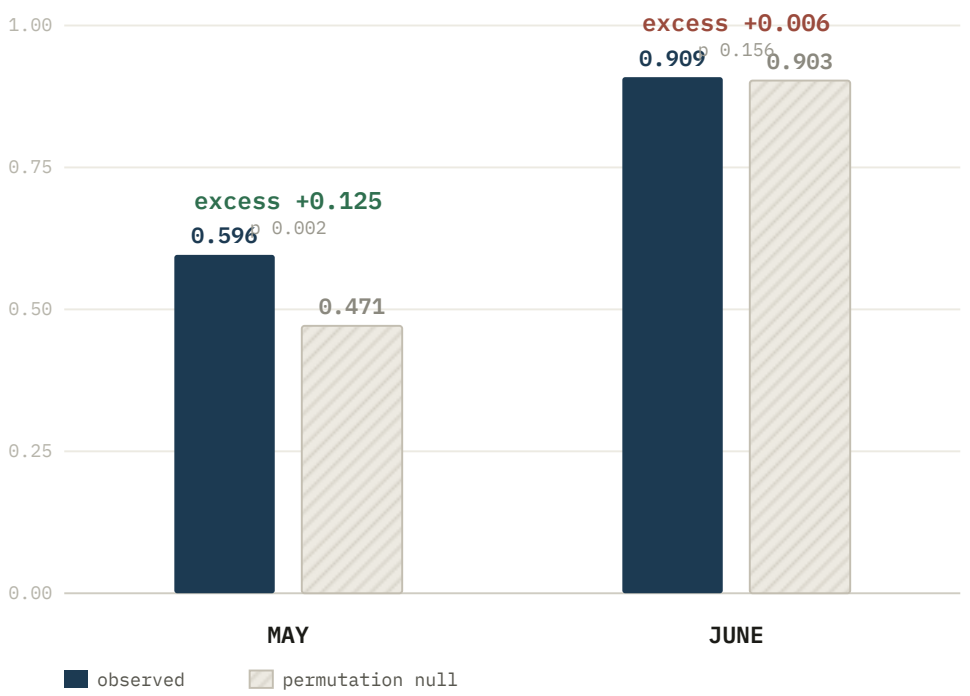
RQ1 · DECISION CONVERGENCE

> June's raw concordance looks like a herding surge: **0.909** against May's **0.596**. The null says otherwise: chance alone predicts **0.903** in June's tape. The apparent convergence is **compositional, not coordinated**.

RQ1 PRIMARY · NOT ESTIMABLE Unchanged. The three-gate cohort is still the single **GPT–Grok** pair; one pair cannot characterize ensemble co-movement. Deferred to Phase B.

The null rose, not the signal

RQ1 EXPLORATORY · ALL-MODEL ACTION CONCORDANCE



	MAY	JUNE
Observed concordance	0.596	0.909
Permutation null mean	0.471	0.903
Excess over chance	**+0.125**	**+0.006**
Permutation p	0.002	0.156

In June's down tape, every model's action mix concentrated (long stretches of holding and paring), so two models agreeing 90% of the time is what independence already predicts. The herding signal collapsed from +12.5 points (p = 0.002) to +0.6 points (p = 0.156), statistically indistinguishable from zero; the 90% CI on observed concordance [0.888, 0.927] straddles the null mean. Portfolio-weight correlation, a complementary lens, sits at 0.49.

WHY THE MONTHS CAN'T BE NAIVELY COMPARED

Raw concordances are computed under **different opportunity sets**.

May ran prompt `regime v1`; June is the first full v2 month. Behavioral metrics are segmented by regime by standing convention, so any May→June delta confounds behavior with the prompt change.

HONEST MONTH-TWO FINDING

Raw concordance is a misleading herding metric on its own; the null-adjusted excess is the estimand. The pilot caught this before the confirmatory phase, which is precisely what a pilot is for.

INCLUSION GATE MATRIX · JUNE

MODEL	COMPLETE	BOOK	IDENTITY	ELIGIBILITY
GPT-5.4	✓ 100%	✓ clean	✓ stable	PRIMARY
Grok 4.20	✓ 100%	✓ clean	✓ stable	PRIMARY
DeepSeek V4	✓ 100%	✓ clean	✗ splice*	EXPLORATORY
Gemini 3.1 Pro	✗ 78.5%	✓ clean	✓ stable	PERF. ONLY
Claude Sonnet 4.6	✓ 99.6%	✗ corrupt	✓ stable	NON-EST.
Claude Opus 4.6	✓ 99.6%	✗ corrupt	✓ stable	NON-EST.

* splice historical (Page 5)

Passing cohort unchanged: **GPT, Grok**

EXCLUSIONS KEPT DISTINCT

Gemini
Incomplete decision record (2nd month).

DeepSeek
Identity splice inside the clean window (historical).

Sonnet · Opus
Corrupt carried-forward books (Phase A).

Model Profiles & Notable Decisions

PROFILES VERBATIM · profiles[] @ a28482d9

GPT-5.4

PRIMARY

Directional conviction, unrewarded this month · de-levered into weakness

EVIDENCE

Trades/day collapsed to 7.5 (May: 33.25 — cross-regime comparison; the v1→v2 prompt change confounds this delta, illustrative only); reversal rate 0.38; cash rebuilt to 6.2% (May: 1.6%); turnover 0.153; avg hold 16.7d; 158 trades; win rate 0.45; monthly −6.34%, max DD −7.0%.

READ

Fewer, larger, longer directional positions under v2; June's losses came from held conviction, not churn.

Grok 4.20

PRIMARY

Two-way trader, defensively effective · fully deployed, fast exits

EVIDENCE

18.9 trades/day (highest of cohort); reversal 0.48; cash 0.5%; RQ2 −0.096; RQ3 −0.052; monthly +0.15% — flat through a −1.55% SPY month.

POSTURE

Fully deployed, fast exits. Best risk outcome of the estimable cohort this month.

DeepSeek V4 ‡

SHOWN ‡

Model-spliced (historical), behavioral exploratory

EVIDENCE

11.0 trades/day; reversal 0.47; cash 2.8%; monthly −0.40%.

No style/posture characterization published: June decisions inherit the historical model-identity splice; per-model RQ2/RQ3 (−0.049 / +0.144) exploratory.

Gemini 3.1 Pro

PERF. ONLY

Performance only · no characterization

EVIDENCE

78.5% completeness; behavioral metrics selection-biased. Executed-tick evidence, not characterized: 7.9 trades/day; reversal 0.69; monthly −0.28%.

Behavioral metrics computed on a selection-biased 78.5% record; not comparable to a full-coverage model.

Claude Sonnet 4.6 · Claude Opus 4.6

"Non-estimable (corrupt Phase-A book); no characterization published." Fuller context sourced from `known_caveats.state_integrity` as in May.

NON-ESTIMABLE

SUPPORTING EVIDENCE METRICS · JUNE

METRIC	GPT	GROK	DEEPSEEK ‡
Trades / active day	7.5	18.9	11.0
Reversal rate	0.38	0.48	0.47
Cash held	6.2%	0.5%	2.8%
Disposition (RQ2)	−0.124	−0.096	−0.049 ‡
Calibration (RQ3)	+0.089	−0.052	+0.144 ‡

‡ DeepSeek RQ2/RQ3 exploratory (identity splice, historical): profile detail, not a reported RQ finding. Six-model pooled RQ2/RQ3 remain superseded (`reported: false`); per-model GPT/Grok primary only.

NOTABLE DECISIONS · TOP TRADES BY VALUE

GPT

BUY UNH $5,998 (6/4) · BUY AMZN $5,848 (6/8) · BUY GOOGL $5,800 (6/9)

Grok

SELL QCOM $9,314 (6/23) · SELL TJX $9,027 (6/25) · SELL SLB $8,606 (6/11)

DeepSeek ‡

BUY ADBE $8,211 (6/17) · SELL MA $7,365 (6/22) · BUY QCOM $7,295 (6/15)

Gemini

SELL HON $11,466 (6/29) · SELL ORCL $9,240 (6/4) · SELL XOM $8,840 (6/2)

Sonnet · Opus

Values reflect corrupt books, not published.

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

Methodology, Data Integrity & RQ Update

SOURCE COMMIT e952c409

MONTH INTEGRITY SUMMARY

The cleanest month of Phase A to date

Single prompt regime (v2, derived from decision-log stamps end-to-end, not config), **no model-identity events, no book-integrity events**, 3 systemic missing ticks (2 on 6/9, 1 on 6/10, all-model scheduler gaps, logged, immaterial). API reliability: GPT 0, Grok 0, DeepSeek 0 failures; Sonnet 1; Opus 1; **Gemini 58 of 270 (21.5%)**, the completeness-gate failure, second consecutive month, evenly spread rather than May's clustered outages; under active diagnosis.

CONSERVATION AUDIT · JUNE CLEAN, FULL MONTH

The share- and cash-conservation audit now covers June end-to-end: 270 clean runs per model across all six models (1,620 model-runs; unit encoded as `per_model_runs`), zero anomalies. 3,100+ runs audited since inception, zero discontinuity events on or after April 23. June is asserted **conservation-audited clean**, the stronger claim May's framework established.

LONG-ONLY BASELINE **553** SELLs · **570** BUYs · **0 SHORT/COVER**

A hard pre-activation baseline for July's shorting before/after comparison, confirmed from the trade record itself.

FIRST MECHANIZED BUILD

The first monthly layer produced by the mechanized report builder rather than hand-walked, under the release gate that blocks emission with a null `source_commit` (this layer: `e952c409`, non-null, self-validated). One line of methods pride, as budgeted.

`data_layer.json @ a28482d9 · source_commit e952c409 · additive-only over 1fc61f0c (17 changes).`

CARRIED INTEGRITY STATE · UNCHANGED

Sonnet/Opus non-estimable for all of Phase A (carried-forward fabrication: ~$297k phantom cash / ~$575k phantom shorts). DeepSeek identity splice historical (4/24 v4-flash → 5/21 repoint, first v4-pro success 5/22): performance real, decision data exploratory. May comparatives cited here are anchor-independent (clean-window returns, RQ values) and unaffected by the SPY anchor correction; the corrected May inception figures live natively in the layer lineage (erratum + rev. 2 handled separately by Reports).

RQ UPDATE · JUNE PASSING COHORT: GPT, GROK

RQ1 · Herding
NOT ESTIMABLE
Single-pair cohort; deferred Phase B. Exploratory concordance 0.909, excess +0.006, p 0.156, not distinguishable from chance.

RQ2 · Disposition
PER-MODEL PRIMARY
GPT −0.124 (PGR 0.022 / PLR 0.146, n=75), Grok −0.096 (PGR 0.032 / PLR 0.128, n=182). Six-model pool superseded.

RQ3 · Calibration
PER-MODEL PRIMARY
GPT +0.089 (n=23 closed, above the 20 minimum), Grok −0.052 (n=80). Six-model pool superseded.

RQ4 · Factor exposure
OPEN
Factor data alignment pending.

RQ5 · Drawdown response
TESTING
Zero drawdown days at the −10% threshold (GPT's −7.0% did not cross); accumulating β +0.0022, p 0.371.

RQ6 · Reproducibility
OPEN
No determinism-probe data yet.

SPY June −1.55%; inception cum +9.75% (anchor 680.40, deployment-moment pin, month-end frozen). Descriptive monthly Sharpe, not deflated; rf convention carried. All figures Phase A pilot / exploratory, non-confirmatory. Shorting activated July 1 and belongs to the July report.