Summary & Leaderboard

REGIME v3 · 22 TRADING DAYS · FIRST SHORTING MONTH

> Shorting went live July 1 with identical prompts and identical caps for all six models.
> **Two used it. Four never shorted once in 31 days** — and the null is the finding.

PRIMARY LEADERBOARD · CLEAN WINDOW (APRIL 23 → JULY 31), DEFAB BASIS

Estimable returns · integrity caveats noted

RANK	MODEL	CLEAN CUM. · 4/23→7/31	VS SPY	STATUS
1	**Gemini 3.1 Pro** Defab-corrected 13.07% → 12.62% (Page 5) · completeness 58.7%, 3rd month	**+12.62%**	+7.58 ▲	PERFORMANCE ONLY
–	**SPY** passive benchmark The line every estimable model is measured against	**+5.04%**	–	BENCHMARK
2	**DeepSeek V4** ‡ Defab-corrected 4.24% → 3.57% (Page 5) · model-spliced, behavioral exploratory	**+3.57%**	−1.47	EXPLORATORY ‡
3	**Grok 4.20** Highest-churn trader in the book, long-only under the new rules	**+3.17%**	−1.87	ESTIMABLE
4	**GPT-5.4** Long-only by choice; the book's longest holder	**+2.14%**	−2.90	ESTIMABLE
HELD OUT · NON-ESTIMABLE, NOT OMITTED FOR RANKING				
–	**Claude Sonnet 4.6** · corrupt book, Phase A (Page 5)	–	–	NON-ESTIMABLE
–	**Claude Opus 4.6** · corrupt book, Phase A (Page 5)	–	–	NON-ESTIMABLE

THE CAPABILITY-UPTAKE NULL · IDENTICAL PROMPTS, IDENTICAL CAPS · THE FOUR ZEROS CARRY EQUAL WEIGHT

20% short cap · 10% short stop · 120% gross ceiling

MODEL	OPENS	COVERS	CLOSED	AVG / MAX EXPOSURE	NOTE
Gemini 3.1 Pro	**36**	26	22	9.25% / 20.09%	Month-end open short: META −8.17% wt, −4.07% unrealized
DeepSeek V4 ‡	**8**	4	2	peak util 0.60	Brief trial; flat since July 14
GPT-5.4	**0**	0	0	–	Never shorted in 31 days of granted capability
Grok 4.20	**0**	0	0	–	Zero SHORT executions, confirmed in the layer
Claude Sonnet 4.6 n-e	**0**	0	0	–	Ran long-only in July
Claude Opus 4.6 n-e	**0**	0	0	–	Ran long-only in July

Heterogeneous uptake of an identical capability under identical instructions is precisely the class of behavioral variation this experiment exists to surface; the four zeros carry equal weight with the activity.

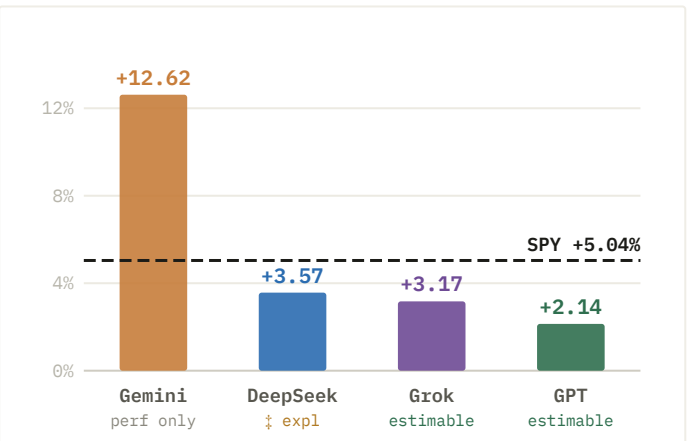
GEMINI'S MONTH IS A TWO-ACT ARC, AND THE RECORD STAYS PLAIN

Best month of the experiment on the book: +6.78% in July against SPY's +0.17%, extending its clean-window lead to +12.62% vs +5.04%. And its third consecutive completeness-gate failure: 58.7% against the 0.80 threshold — it fails as ruled, no softening. What changed in July is that the cause is now diagnosed, fixed, and verified in situ: hidden thinking was spending against a 4,096-token shared output budget, truncating decisions mid-string with `finish_reason=MAX_TOKENS` — the mechanism behind the May 0.57 / June 0.79 / July 0.51 (pre-fix) failures. The budget was equalized to 16,384 effective 7/28; the layer segments the month at that boundary: **`51.1% pre-fix, 92.3% post-fix`**. Verification: 0 MAX_TOKENS in 39 post-fix cycles; the two residual 7/29 failures are the pre-existing, budget-independent STOP-class parse mode. The Phase B validation clock — ≥0.80 completeness and ≤1% MAX_TOKENS each month at the deployed configuration, identity stable, no averaging — runs through October 31; following an SDK integration-path migration landed August 2 (ledger-documented, config-semantics-preserving), the clock starts at the first post-migration decision success, with the August segment counted over its ~4 weeks.

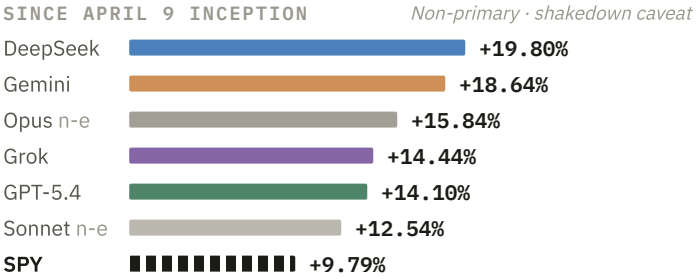
JULY MONTHLY RETURN

Gemini	Opus n-e	Sonnet n-e	SPY	GPT	DeepSeek	Grok
+6.78%	+2.10%	+2.04%	+0.17%	+0.16%	−0.89%	−1.51%

CLEAN-WINDOW CUMULATIVE · ANCHORED 4/23



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



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; see Data Integrity. SPY anchor 680.40.

Performance & Risk Detail

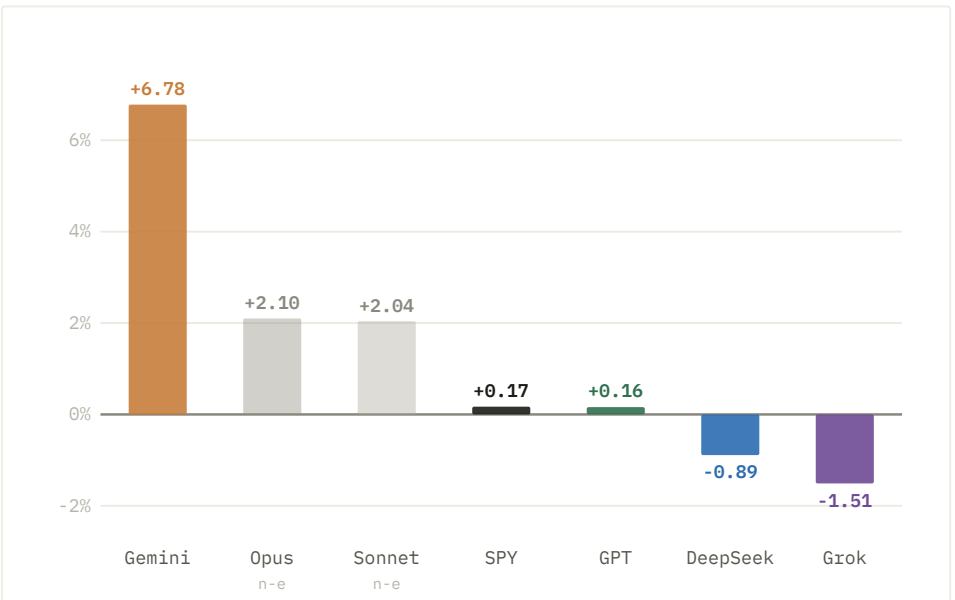
DESCRIPTIVE SHARPE · NOT DEFLATED

> In a flat tape, **the book that took the most novel risk posted the best risk-adjusted month** — and the first negative short P&L of the experiment sits inside it.

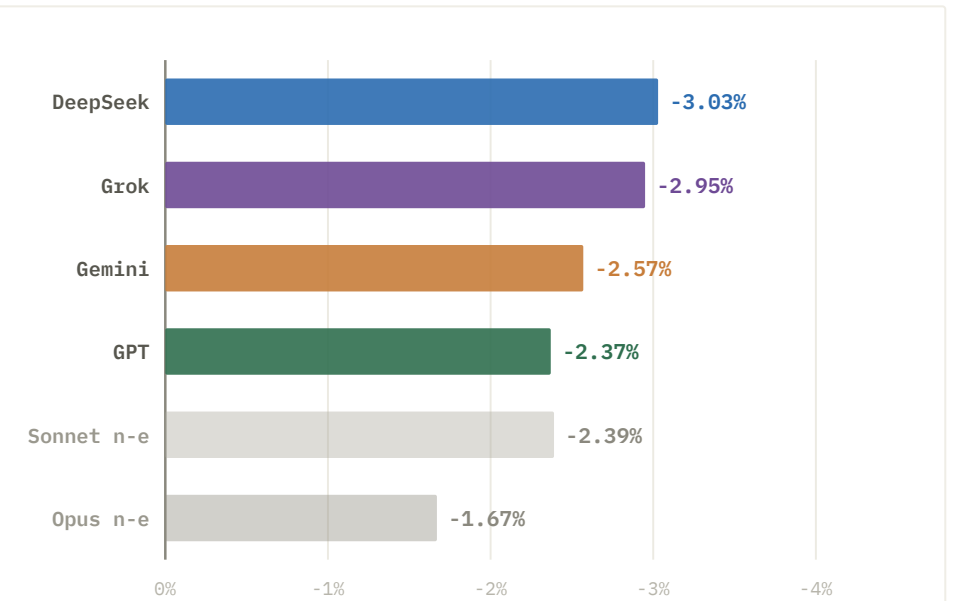
RISK METRICS · JULY CALENDAR MONTH

MODEL	DESCRIPTIVE SHARPE	VOLATILITY (ANN.)	MAX DD (EOD)	STATUS
Gemini 3.1 Pro · best risk-adjusted month; heaviest shorter	**5.10**	0.149	−2.57%	PERFORMANCE ONLY
Claude Opus 4.6 · off corrupt carried-forward book	2.15	0.101	−1.67%	NON-ESTIMABLE
Claude Sonnet 4.6 · off corrupt carried-forward book	1.73	0.123	−2.39%	NON-ESTIMABLE
SPY benchmark · essentially flat	**−0.07**	–	–	BENCHMARK
GPT-5.4 · a flat month traded almost perfectly flat	**−0.13**	0.100	−2.37%	ESTIMABLE
DeepSeek V4 ‡ · clean book; splice historical	**−1.45**	0.096	−3.03%	EXPLORATORY ‡
Grok 4.20 · worst estimable July	**−1.77**	0.120	−2.95%	ESTIMABLE

JULY MONTHLY RETURN · BY BOOK



MAX INTRA-MONTH DRAWDOWN (EOD) Clustered −1.7% to −3.0% · no halt approached



SHORTING RISK DETAIL · FIRST MONTH

Gemini's gross exposure averaged 107.0% of equity (max 118.9%, within the 120% ceiling); net exposure averaged 88.0%. Its maximum short-utilization-vs-cap printed **1.0045** — not a breach: the 20% cap is a fill-time constraint (enforced at fill in the risk engine), not a continuously-maintained invariant, so a book filled to 99.88% of cap crosses it on any adverse mark with no order involved. The 1.0045 observation sits on a tick with no position-moving execution — the last fill before it cleared at 0.9988× cap and the next tick covered to 0.1699. Verified across all July ticks: zero fills cleared above the cap, for any model. At month-end it carried the cohort's only open short: META, −17.5522 shares, −8.17% of equity, −4.07% unrealized. DeepSeek's peak utilization was 0.60. The four non-shorters ran gross ≈ net ≈ 96–99% long throughout.

BASIS CAVEATS: GPT, Gemini & Grok daily risk metrics are raw-basis-sensitive (layer `risk_basis` flags); DeepSeek's span the splice; Sonnet/Opus excluded, corrupt books.

SHORT-SIDE RESULTS · EXPLORATORY, LOW-N, FLAGGED HARD

0.292	0.435	−0.157
short hit-rate (24 closed)	long hit-rate (92 closed)	conf-outcome corr, short (long +0.223)

First-month shorts were **anti-calibrated**: stated confidence inversely predicted short outcomes in month one. Gemini's realized short P&L −$539 (hit 0.318, 7 of 22); DeepSeek's two closed shorts both lost (−$726 total). Single month, n=24, side-segmented replay outside the canonical RQ3 pipeline — exploratory, no trend claim. It is nonetheless the most interesting behavioral line of the month: the first capability extension produced immediate, measurable overconfidence on the new side of the book.

Cross-Model Behavioral & RQ1 Herding

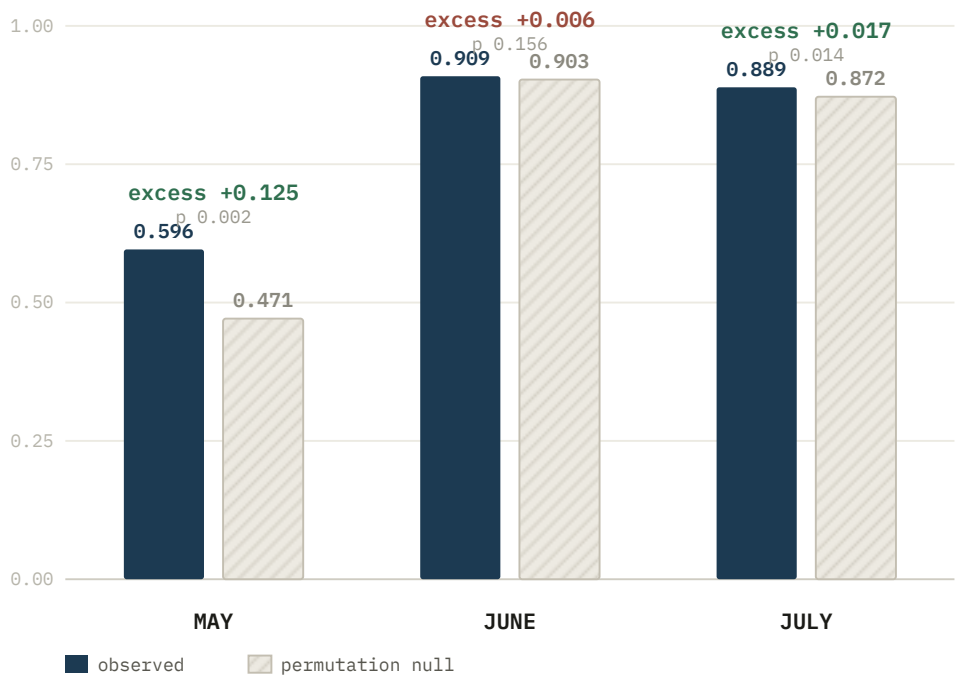
RQ1 · DECISION CONVERGENCE

> July shows the **first statistically distinguishable herding excess since May** — small, real, and reported the way June taught us: all three numbers, never the observed level alone.

RQ1 PRIMARY · NOT ESTIMABLE Unchanged. Single-pair three-gate cohort (**GPT, Grok**); deferred to Phase B.

A real but small excess

RQ1 EXPLORATORY · ALL-MODEL ACTION CONCORDANCE



	MAY	JUNE	JULY
Observed concordance	0.596	0.909	0.889
Permutation null mean	0.471	0.903	0.872
Excess over chance	**+0.125**	**+0.006**	**+0.017**
Permutation p	0.002	0.156	**0.014**

July: 0.889 observed (n = 185 pairwise, 283 ticks) against a 0.872 null: excess +0.017, p = 0.014, CI [0.853, 0.915]. A single market regime end-to-end — range-bound on 22 of 22 classified days — so no within-month regime mixing dilutes the read. The high absolute level is HOLD-share; chance alone predicts 87% agreement in this tape, which is exactly why excess-over-null is the estimand. On that estimand July is the first month since May with a distinguishable excess: real, and roughly one-seventh of May's +0.125. Weight correlation 0.486, essentially unchanged from June. Levels are not comparable across columns (v1/v2/v3 prompt regimes); only each month's excess against its own null. Exploratory, non-confirmatory.

INCLUSION GATE MATRIX · JULY

MODEL	COMPLETE	BOOK	IDENTITY	ELIGIBILITY
GPT-5.4	✓ 100%	✓ clean	✓ stable	PRIMARY
Grok 4.20	✓ 100%	✓ clean	✓ stable	PRIMARY
DeepSeek V4	✓ 99.3%	✓ clean	✗ splice*	EXPLORATORY
Gemini 3.1 Pro	✗ 58.7%	✓ clean	✓ stable	PERF. ONLY
Claude Sonnet 4.6	✓ 100%	✗ corrupt	✓ stable	NON-EST.
Claude Opus 4.6	✓ 100%	✗ corrupt	✓ stable	NON-EST.

* splice historical · first-ever API failures: 2, transport-level, gate unaffected

Passing cohort unchanged: **GPT, Grok**

DEEPSEEK READS THROUGH THE COLLISION LENS

Its July record spans the 7/20 V4-GA boundary, a **disclosed gap**: no pre-launch snapshot exists, so a build swap under the alias is unobservable by design. What is observable is clean: the alias echo was unchanged across the boundary (231/231 July observations returned `deepseek-v4-pro`), with zero behavioral/failure/latency shift either side of the GA date. Its shorting before/after must not be read in isolation from this boundary.

RQ2 / RQ3 · PER-MODEL PRIMARY (POOLS SUPERSEDED)

RQ2 disposition	GPT −0.099 · Grok −0.129
RQ3 calibration	GPT +0.395 · Grok −0.013

GPT's jump from June's +0.089 is noted neutrally: n=26, single month, no trend claim.

Model Profiles & Notable Decisions

PROFILES VERBATIM · profiles[] @ FINAL LAYER

GPT-5.4

PRIMARY

Long-only by choice; the book's longest holder · patient, near-flat, unshorted

EVIDENCE

Never shorted (0 opens in 31 days of granted capability); avg hold extended again to 22.9d (June: 16.7d — cross-regime, illustrative only); 10.4 trades/day; reversal 0.37; cash 4.0%; turnover 0.189; 26 closed trades, day win rate 0.52; monthly +0.16% in a +0.17% SPY month; RQ3 +0.395 (n=26, no trend claim).

READ

Declined the new capability entirely and slowed further into a holder's book; a flat month traded almost perfectly flat.

Grok 4.20

PRIMARY

Highest-churn trader in the book, long-only under the new rules · fully deployed, unhedged

EVIDENCE

Never shorted (0 opens); 20.8 trades/day and turnover 0.497, both cohort-highs; cash 0.9%; hold 7.0d; 85 closed trades, day win rate 0.57; monthly −1.51%, worst estimable July; RQ2 −0.129, RQ3 −0.013.

READ

Kept maximum activity and full deployment through a flat tape and paid for it; a high daily win rate lost to asymmetric losers.

DeepSeek V4 ‡

SHOWN ‡

Model-spliced (historical), behavioral exploratory

EVIDENCE

Brief shorting trial: 8 opens / 4 covers, 2 closed shorts, both losses (−$726); peak utilization-vs-cap 0.60; 8.9 trades/day; hold 21.1d; cash 6.1%; monthly −0.89%. July spans the 7/20 V4-GA disclosed-gap boundary (alias echo 231/231 unchanged, zero observed shift) — shorting before/after reads through the collision lens, per ledger.

Gemini 3.1 Pro

PERF. ONLY

Performance only · no characterization; segmentation and shorting activity are evidence, not style

EVIDENCE

Completeness 58.7% — third consecutive gate failure — segmented at the 7/28 budget equalization: 51.1% pre-fix / 92.3% post-fix, 0 MAX_TOKENS in 39 post-fix cycles. The cohort's only heavy shorter: 36 opens / 26 covers, 22 closed (hit 0.318, −$539 realized), avg/max short exposure 9.25%/20.09%, max utilization-vs-cap 1.0045 (at the cap; enforcement at fill). Monthly +6.78%, best of the experiment; behavioral metrics selection-biased, not characterized.

Claude Sonnet 4.6 · Claude Opus 4.6

"Non-estimable (corrupt Phase-A book); no characterization published." Both books ran long-only in July (0 short opens each — layer short-activity block).

NON-ESTIMABLE

SUPPORTING EVIDENCE METRICS · JULY

METRIC	GPT	GROK	DEEPSEEK ‡
Trades / active day	10.4	20.8	8.9
Reversal rate	0.37	0.44	0.42
Cash held	4.0%	0.9%	6.1%
Avg hold (days)	22.9	7.0	21.1
Disposition (RQ2)	−0.099	−0.129	−0.072 ‡
Calibration (RQ3)	+0.395	−0.013	+0.277 ‡

‡ exploratory (identity splice, historical): profile detail, not a reported RQ finding.

NOTABLE DECISIONS · TOP TRADES BY VALUE

GPT

BUY LLY $6,868 (7/23) · BUY UNP $5,762 (7/24) · BUY AMD $4,684 (7/28)

Grok

SELL UNP $9,091 (7/29) · SELL ISRG $7,905 (7/14) · SELL AVGO $7,244 (7/23)

DeepSeek ‡

SELL DE $9,426 (7/7) · COVER QCOM $9,412 (7/9) · SELL RTX $8,547 (7/10)

Gemini

SELL USO $20,837 (7/30) · COVER PG $16,505 (7/30) · SELL ABT $14,809 (7/29)

Sonnet · Opus

Values reflect corrupt books, not published.

Correction (rev. 2): the notable-decisions extractor omitted short-side executions (SHORT/COVER) from July 1 until corrected August 10 — a lossy BUY/SELL filter across 74 short-side orders, ledgered as `notable_events_long_only_extractor`. The list now includes all execution types: Gemini's PG cover and DeepSeek's QCOM cover enter at #2, displacing SLB $13,020 and GLD $7,903. Rev. 1's "long-leg only by extractor design" framing described a defect as a design property; this revision states it as what it was — an omission, now corrected. Short-side aggregates remain on Pages 1–2.

Methodology, Data Integrity & RQ Update

SOURCE COMMIT cf8c0488

MONTH INTEGRITY SUMMARY

Single regime end-to-end, one systemic event, one pre-publication correction on the record

Prompt regime v3 end-to-end: **1,698 of 1,698 decision records stamp v3** (283 × 6, the layer's authoritative boundary). Market regime range-bound on 22/22 classified days. One systemic event: 7/24, 3 of 13 cycles lost cohort-wide (12:30, 14:00, 14:30 ET; chain/scheduler cause, all six equally), self-excluded, denominator 283 of 286. API reliability: GPT 0, Grok 0, Sonnet 0, Opus 0; DeepSeek 2 (transport-level, its first ever, 99.3%, gate unaffected); **Gemini 117 (58.7%)**, the gate failure (Page 1).

PRE-PUBLICATION CORRECTION · DEFAB DIRECTION-FIX, ON THE RECORD

provenance.corrections.defab_short_blind_replay

The de-fabrication replay was direction-blind (BUY/SELL fills only), so the two shorted books' clean-window returns excluded all short-side cash and P&L: `Gemini 13.07% → 12.62%` · `DeepSeek 4.24% → 3.57%`; rank order unchanged; long-only models and SPY untouched. The defect was caught by the module's own overhang-constant conservation check: DeepSeek's clean-window cash drift, −$726.2341, equals its realized short P&L to the cent, and Gemini's +$9,213.10 drift carried 17.5522 shares of untracked short inventory (independently corroborating the month-end META open short). Committed-then-corrected pre-publication, same class as the SPY inception-anchor correction. May replays byte-identical, June exact — the fix is a no-op on long-only months, as it must be.

CONSERVATION AUDIT · JULY REGISTERED

283 clean runs per model across all six (unit `per_model_runs`), zero anomalies; 6,000+ runs audited since inception, zero discontinuity events on or after April 23. July is conservation-audited clean — the first audited month containing short positions, extending the conservation replay across two-sided books.

THREE LEDGER BOUNDARIES RIDE IN THIS LAYER

7/1 · Shorting activation (prompt v3)
20% short-exposure cap, 10% short stop, 120% gross ceiling. The trade record confirms June's long-only baseline held until activation.

7/20 · DeepSeek V4-GA — disclosed gap
No pre-launch snapshot; alias echo unchanged 231/231; legacy aliases (`deepseek-chat`, `deepseek-reasoner`) hard-stopped 7/24 with zero errors (the lab moved off `deepseek-reasoner` 5/21); zero observed behavioral/failure/latency shift; the capture mechanism records only the echoed alias, so a build swap under the alias is unobservable by design. Disclosed with the same candor as the first splice, as the layer itself requires.

7/28 · Gemini budget equalization
`max_output_tokens` 4096 → 16384 (landed 7/27, effective first tick 7/28, per the 7/27 Research ruling). Infrastructure class: no monthly change slot consumed, no version boundary, identity unchanged — mirrors the 5/21 DeepSeek raise. Evidence: production-replica probe (arm A: 22/23 failures MAX_TOKENS at 4096; arm D: 15/15 STOP at 16384), in-situ verification 7/28–7/30 (39/39 STOP, 0 MAX_TOKENS), residual 7/29 STOP-class parse mode documented, daily completeness floor never breached post-fix. Time-series analyses spanning 7/28 must treat it as a regime point. Phase B validation clock, per the 8/2 superseding ledger entry: from the first logged post-migration decision success through 2026-10-31 — an SDK integration-path migration (`google-generativeai` → `google-genai`, wire-verified as configuration-preserving, staged on main 8/2, cutover recorded at observation) restarts the clock at that first post-migration success; each month ≥0.80 completeness AND ≤1% MAX_TOKENS at this configuration AND identity stable, no averaging; the August segment counts as a clean-month observation over its ~4 weeks.

RQ UPDATE · JULY PASSING COHORT: GPT, GROK

RQ1 · Herding
NOT ESTIMABLE · EXPL. SIGNAL
Primary deferred to Phase B (single pair). Exploratory: excess +0.017 over a 0.872 null, p 0.014 — first distinguishable excess since May; presented per the June null-shift lesson.

RQ2 · Disposition
PER-MODEL PRIMARY
GPT −0.099 (n=109 sales), Grok −0.129 (n=210). Pool superseded, not reported.

RQ3 · Calibration
PER-MODEL PRIMARY
GPT +0.395 (n=26, neutral note), Grok −0.013 (n=85). Short-side calibration (Page 2) exploratory, outside the canonical pipeline.

RQ4 · Factor exposure
OPEN

RQ5 · Drawdown response
TESTING
No −10% crossings; GPT/Grok drawdowns ≤3%.

RQ6 · Reproducibility
OPEN

SPY July +0.17%; inception cum +9.79% (anchor 680.40). Descriptive monthly Sharpe, not deflated; rf convention carried. All figures Phase A pilot / exploratory, non-confirmatory. Statuses carried: Sonnet/Opus non-estimable (Phase A), DeepSeek exploratory (splice, historical). Gemini's Phase B eligibility rides the Aug–Oct validation clock.