

Weekly Model Watch #4

September 2, 2026 · MarketMania Research · Weekly series · window Aug 24-30, 2026 UTC

RESEARCH SNAPSHOT

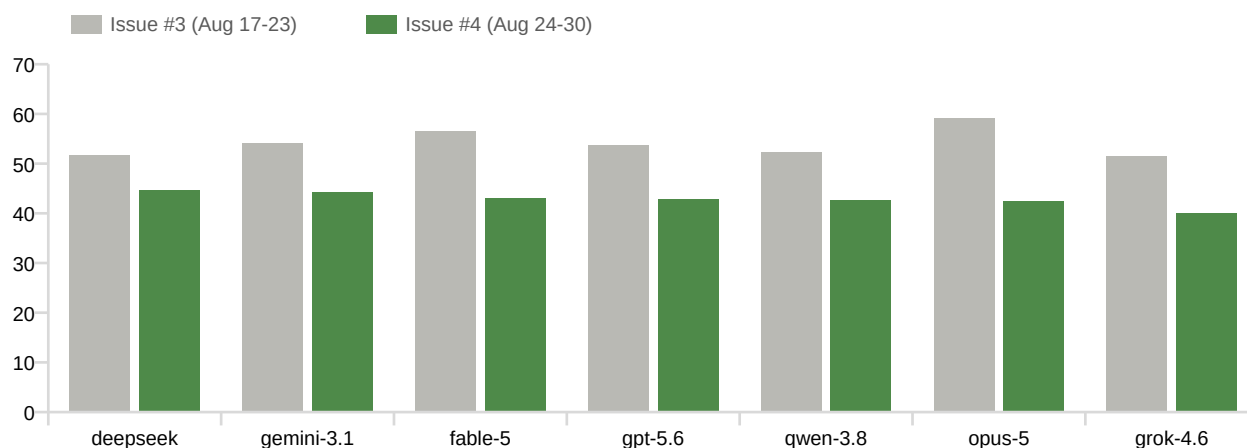
4,576 directional calls scored	7 models (grok line spliced)	of 9,301 mature forecasts	window Aug 24-30, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 42.9%	cutoff Mon Aug 31, 2026, 16:00 UTC
market: BTC net -0.07% (prior +23.58%)	ann vol 30.9% (was 71.1%)	TOP5 volume \$19.72B , -24% w/w	pairwise corr 0.85 (was 0.81)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 54.0% -> 42.9% (-11.1pp) with 0 of 7 models improving, and no weekly title was awarded, for the fourth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** deepseek-v4-pro leads descriptively at 44.6% [41.0%, 48.2%]; the gap to #2 (gemini-3.1-pro, 44.2%) is 0.3pp. 0 of 7 models improved week over week; issue #3's leader claude-opus-5 is now 6 of 7.
- **No ticker cleared 51%.** SOL is the week's most readable coin (45.8%, n=1,025) and XRP the hardest at 40.3%; 0 of the 5 assets cleared 51%. Pair of the week: qwen-3.8-max x SOL 48.7% (n=150); worst: claude-opus-5 x XRP 34.6% (n=104).
- **3 of 4 qualifying turns found a caller.** BTC short -2.49%, 8 caller(s), first gpt-5.6-sol; ETH long +2.31%, 13 caller(s), first gpt-5.6-sol; ETH short -2.80%, 0 caller(s), first NOBODY; ETH short -2.52%, 1 caller(s), first deepseek-v4-pro.
- **Self-agreement read as a penalty in every cell.** Cross-TF/FH lifts ran -7.6pp to -1.9pp (issue #3: -0.6 to +4.6pp). On trend days, 4h TF-agreement hit 41.9% (n=248) against 36.0% (n=225) on flat days.



Directional hit-rate (%) by model, week over week, ranked by this week's rate. Short names: opus-5 = claude-opus-5, fable-5 = claude-fable-5, gemini-3.1 = gemini-3.1-pro, gpt-5.6 = gpt-5.6-sol, qwen-3.8 = qwen-3.8-max, deepseek = deepseek-v4-pro. Field base 54.0% -> 42.9%. The grok column compares the spliced 4.5+4.6 line against issue #3's pure grok-4.5. "Was" values are the numbers published in issue #3; the leaderboard's W/w column is computed on unrounded rates and can differ by 0.1pp from the difference of the rounded columns.

Why it matters

MarketMania scores 7 model lines against the same market, hour after hour. Weekly Model Watch asks four practical questions about that week of calls: who led the week, which tickers were readable, who saw the turn first, and does a model agreeing with itself across timeframes/horizons make its call stronger? Each question gets its own block below, with its own table, its own N, and its own honesty caveats. Issue #4 is the first issue whose grok row is a lineage splice, so read that row as a line, not a single model build.

How to read this

Every block below is scored on the same pool: matured directional calls inside the 1h/4h/1d FH gate, hit by the methodology v1.1 direction rule. The weekly title is awarded only on a ≥ 5 pp gap AND non-overlapping 95% Wilson CIs with $N \geq 10$ per side -- a descriptive lead is not a title. Reversals use the 4h grid with a $\geq 2.0\%$ counter-move sustained 8h, BTC/ETH only in v1. Cross-TF/FH blocks ask whether a model agreeing with itself does any better, and are split trend vs flat wherever the week supplies both sides. The grok row is the lineage splice grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22).

Leaderboard of the week

Model	N	Cov.	Hit rate	95% CI	W/w pp
deepseek-v4-pro	729	54.8%	44.6%	41.0%-48.2%	-7.0
gemini-3.1-pro	755	56.9%	44.2%	40.7%-47.8%	-9.9
claude-fable-5	651	49.2%	43.0%	39.3%-46.8%	-13.5
gpt-5.6-sol	729	54.8%	42.7%	39.1%-46.3%	-11.0
qwen-3.8-max	672	50.5%	42.6%	38.9%-46.3%	-9.7
claude-opus-5	561	42.2%	42.4%	38.4%-46.6%	-16.7
grok-4.6 (spliced)	479	36.0%	39.9%	35.6%-44.3%	-11.5

No weekly title is awarded this issue -- for the fourth time. Title rule: ≥ 5 pp gap AND non-overlapping 95% CI, $N \geq 10$ per side. deepseek-v4-pro's lead over gemini-3.1-pro is 0.3pp and their CIs overlap, so neither condition is met. grok-4.6 (spliced) = grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22).

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #3
SOL	1,025	45.8%	42.7%-48.8%	54.3%
BTC	940	43.7%	40.6%-46.9%	57.5%
ETH	827	43.3%	40.0%-46.7%	52.8%
BNB	895	41.2%	38.0%-44.5%	54.0%
XRP	889	40.3%	37.1%-43.5%	51.2%

SOL was the field's most readable ticker (45.8%) and XRP the hardest (40.3%); 0 of 5 assets cleared 51% this week (issue #3: 5 of 5). Issue #3's best ticker was BTC and its hardest XRP -- this issue's ordering is SOL > BTC > ETH > BNB > XRP.

Pair of the week

Top 6 of the 12 model x ticker pairs tracked this week (best hit-rate), then the 5 worst in their own table -- the pair list is split rather than paged so neither table breaks across a page. Pairs shown require N of 8 or more calls.

Model (top 6)	Symbol	N	Hit rate
qwen-3.8-max	SOL	150	48.7%
gemini-3.1-pro	ETH	130	47.7%
claude-opus-5	SOL	131	47.3%
deepseek-v4-pro	SOL	154	46.1%
gemini-3.1-pro	XRP	154	46.1%
gpt-5.6-sol	SOL	157	45.9%

Model (5 worst)	Symbol	N	Hit rate
grok-4.6 (spliced)	ETH	81	38.3%
claude-fable-5	XRP	116	37.1%
qwen-3.8-max	XRP	135	36.3%
grok-4.6 (spliced)	BNB	100	35.0%
claude-opus-5	XRP	104	34.6%

qwen-3.8-max x SOL (48.7%, n=150) is the week's best pair and claude-opus-5 x XRP (34.6%, n=104) the weakest. Pair history accumulates across issues -- read this as the fourth data point, not a ranking.

Who saw the reversal first

Reversal definition: 4h grid; trend = sign of prior 24h; counter-move of 2.0% or more (BTC/ETH) sustained 8h. A model 'calls' the reversal if it has a matured directional hit call in the new direction, FH 4h or 1d, in the slot window [T-12h, T+FH].

Symbol	Time (UTC)	New side	Move (8h)	Callers	First caller
BTC	Fri Aug 28, 12:00	Short	-2.49%	8	gpt-5.6-sol
ETH	Wed Aug 26, 16:00	Long	+2.31%	13	gpt-5.6-sol
ETH	Fri Aug 28, 12:00	Short	-2.80%	0	NOBODY
ETH	Sun Aug 30, 16:00	Short	-2.52%	1	deepseek-v4-pro

The BTC short turn drew 8 caller(s), first gpt-5.6-sol on a 4h/1h call at Aug 28, 04:01 with confidence 62; the ETH long turn drew 13 caller(s), first gpt-5.6-sol on a 1d/4h call at Aug 27, 00:01 with confidence 69; the ETH short turn drew NOBODY; the ETH short turn drew 1 caller(s), first deepseek-v4-pro on a 4h/4h call at Aug 30, 04:01 with confidence 56. Scope in v1 remains BTC/ETH only; the event count across four issues is still far too small to characterise turn-detection skill either way.

Cross-TF and cross-FH confirmation: does agreeing with yourself help?

Block	FH	vs.	Agree n	Agree hit	Dis n	Dis hit	Base hit	Lift
cross-TF	4h	TF 4h vs 1h	473	39.1%	18	11.1%	41.0% (712)	-1.9
cross-TF	1d	TF 1d vs 4h	105	49.5%	10	100.0%	55.9% (143)	-6.4
cross-FH	4h	by FH 1h	429	38.9%	25	24.0%	41.0% (712)	-2.1
cross-FH	1d	by FH 4h	89	48.3%	10	90.0%	55.9% (143)	-7.6

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 218 / 28 / 258 / 44. Pooled across all four cell families the lifts run -7.6pp to -1.9pp this week (issue #3: -0.6 to +4.6pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has 3 trend days of 7: 08-24 +1.61% (trend), 08-25 -0.58%, 08-26 +0.63%, 08-27 +1.53% (trend), 08-28 -2.91% (trend), 08-29 +0.48%, 08-30 -0.70%. On trend days, 4h TF-agreement hit 41.9% (n=248) against 36.0% (n=225) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,323	34.8%	14.4%	50.8%	2.41
claude-opus-5	1,330	29.2%	13.0%	57.8%	2.24
deepseek-v4-pro	1,330	32.3%	22.5%	45.2%	1.44
gemini-3.1-pro	1,328	35.2%	21.6%	43.1%	1.63
gpt-5.6-sol	1,330	30.3%	24.5%	45.2%	1.24
grok-4.6 (spliced)	1,330	18.3%	17.7%	64.0%	1.03
qwen-3.8-max	1,330	31.6%	18.9%	49.5%	1.67

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 1.66 this week. 'Sideways' was 50.8% of mature forecasts (4,725 of 9,301). Consensus skew: 539 symbol/FH/TF cells had 5 or more models on the same side and hit 40.0% (n=3,371) against the 42.9% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	SOL	Long	85	4h / 4h	Aug 24, 16:01	EXPIRY	-\$0.28

Fail of the week. The week's single highest-confidence individual miss, exit reason expiry. Listed as a single card, never as a model ranking.

Calibration bridge

deepseek-v4-pro is both this week's hit-rate leader (44.6%) and the best-calibrated line (Brier 0.2760, gap +15.9pp). Per-model ok-rate this window ran 99.52% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #4.

Market check: the tape cooled off

OBSERVATION — mean |1d move| 4.28% -> 2.35% (-45% rel), BTC realized vol 58.7% -> 37.9% (ann., hourly); field directional accuracy 54.0% -> 42.9% (-11.1pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL +\$889.37 -> -\$473.59 (adjacent calendar weeks Aug 17-23 vs Aug 24-30; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 53.5% -> 44.2% — the same direction as the trade-based hit rule.

Measure	Aug 17-23 (week A)	Aug 24-30 (week B)	Change
Mean 1d move (5 assets)	4.28%	2.35%	-45% rel
BTC realized vol (ann., hourly)	58.7%	37.9%	-20.8pp
Field directional accuracy	54.0%	42.9%	-11.1pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	53.5%	44.2%	-9.3pp
Field sim win-rate	46.3%	37.0%	-9.2pp
Field sim net PnL	+\$889.37	-\$473.59	--

Week A is exactly the issue-#3 window, so this slice and every week-over-week column in this report compare the same two back-to-back weeks. Descriptive, one pair of weeks; the hit rule is methodology v1.1 (trade-based) and the raw price-sign row is the robustness check (hourly closes at :00 against slots at :01). BTC realized vol here is the hourly-return series from the alive slice; the snapshot row's figure is the audited daily-candle estimate -- different estimators, both reported as measured. Lineage splice ACTIVE: grok-4.5 rows are aggregated into grok-4.6 (flip 2026-08-24T09:22:00+00:00, inside week B). Week A (2026-08-17..2026-08-23) is pure grok-4.5 and reproduces the published issue #3; week B mixes the 4.5-era (Aug 24 00:00-09:22) with 4.6.

Practical implications

- Do not read this week's leaderboard as a skill ranking: no title was awarded and the top CIs overlap.
- Treat per-ticker readability as weekly weather. Four issues, and the best ticker has changed hands every time (SOL -> BNB -> BTC -> SOL).
- Self-agreement lifts run -7.6pp to -1.9pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Aug 24-30, 2026 UTC). Descriptive for this window only; no claim about next week or any model's underlying skill. This window ran 3 trend days of 7 (issue #3: 5 of 7); every comparison with issue #3 is a comparison across regimes as well as across weeks.
- Observations inside one window are not independent, and the 95% Wilson CIs shown throughout are descriptive, not inferential.
- The reversal detector remains v1 and BTC/ETH-only, on price series reconstructed from trade entry prices. 8 events across four issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=18, 10, 25, 10); cells under N=10 are insufficient and never used to rank models. Week-over-week deltas are computed on unrounded rates and may differ by 0.1pp from the difference of the rounded columns.

- Model lines in this window: claude-fable-5, claude-opus-5, deepseek-v4-pro, gemini-3.1-pro, gpt-5.6-sol, grok-4.6 (spliced), qwen-3.8-max. Series density and lineage notes (wave 4): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 24-30) daily coverage is FULL for the first time in the series: the 1w series has slots on 7 of 7 days and the 1M series on 7 of 7 days; both sit outside this report's FH gate (1h/4h/1d) and appear only in the exclusions counter (1w 488, 1M 490). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, INSIDE this window: every grok row in this issue is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)" (110 rows from the 4.5 era, 1,360 from 4.6); issue #3 was pure grok-4.5, so the grok week-over-week row joins a spliced line to a single-model line. (3) qwen-3.8-max succeeded qwen-3.7-max on Aug 5, 2026 (fully before this window).
- Market-state row is computed on a single-exchange (binance) daily candle series, as in issue #3 after the audit that found the raw table mixes two exchanges; issue #3 values are as published.
- Research-to-date counter: directional forecasts resolved inside the FH gate (1h/4h/1d) since Jul 11, counted at the issue cutoff (Aug 31 16:00 UTC) -- the definition pinned in issue #3, which printed 33,809 at the Aug 24 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 38,385 is an additive step under one definition; counter deltas against issue #2 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,301
OK in gate	9,301	-- of them directional	4,576
Out of gate (1w / 1M)	488 / 490	-- of them sideways	4,725
Invalid	11	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-08-24.json	Report cutoff	Mon Aug 31, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-02 06:14 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,576 scored observations. MARKETMANIA RESEARCH TO DATE (as of Aug 31, 2026 cutoff): 38,385 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 17 published reports.

Issue #4. Weekly Model Watch is a living series. Issue #3's open questions -- does the leaderboard reshuffle again, does the reversal detector find callers on a non-ETH turn, and does the every-ticker-above-51% floor survive? -- read this issue as: deepseek-v4-pro at the top (44.6%), 3 of 4 qualifying turns found a caller, and 0 of 5 tickers above 51%. The grok row is a lineage splice this issue (grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)) because the 4.5 -> 4.6 flip happened inside the window. Engine 1.1 has powered the sandbox since Aug 18, i.e. before this window; no cross-engine PnL comparisons are claimed.

What we're testing next

- Next issue: does any model hold the top rank two issues running, and does the title rule ever fire?
- Does the reversal detector find callers outside ETH?
- Monthly series: cross-TF/FH lifts with regime control across mixed weeks (trend vs. flat), block bootstrap.

Related research

Report	Direct PDF link
Consensus Watch #4	https://marketmania.ai/research/reports/consensus-watch-2026-08-24.pdf
Weekly Calibration #4	https://marketmania.ai/research/reports/weekly-calibration-2026-08-24.pdf
Weekly Model Watch #3	https://marketmania.ai/research/reports/model-watch-2026-08-17.pdf

The three weekly reports publish together as one issue each week; each links straight to the others' PDF and to its own previous issue. Direct links are the posting rule from wave 2 on.

CITE THIS REPORT

```
@misc{mm_modelwatch_2026w35,
  title = {Weekly Model Watch #4: weekly leaderboard, ticker/pair reads and cross-confirmation, Aug
24-30 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {September}, day = {2},
  url   = {https://marketmania.ai/research/reports/model-watch-2026-08-24.pdf},
  note  = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Aug 24-30, 2026 UTC; source
weekly_metrics_2026-08-24.json}
}
```