

Weekly Model Watch #5

September 8, 2026 · MarketMania Research · Weekly series · window Aug 31-Sep 6, 2026 UTC

RESEARCH SNAPSHOT

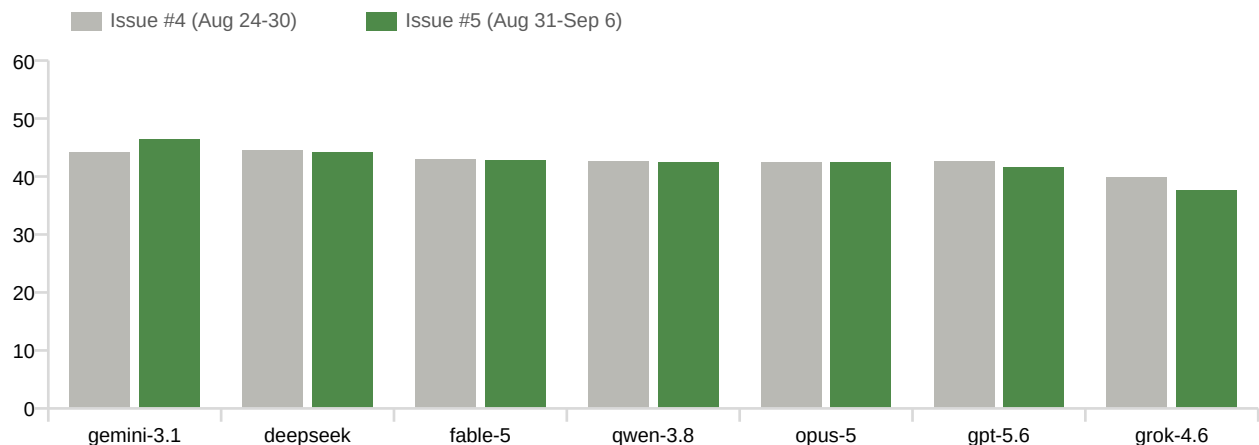
4,520 directional calls scored	7 models (stable lineup)	of 9,264 mature forecasts	window Aug 31-Sep 6, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 42.8%	cutoff Mon Sep 7, 2026, 16:00 UTC
market: BTC net +3.42% (prior -0.07%)	ann vol 43.4% (was 30.9%)	TOP5 volume \$16.24B , -18% w/w	pairwise corr 0.78 (was 0.85)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 42.9% -> 42.8% (-0.1pp) with 1 of 7 models improving, and no weekly title was awarded, for the fifth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** gemini-3.1-pro leads descriptively at 46.5% [43.0%, 50.1%]; the gap to #2 (deepseek-v4-pro, 44.2%) is 2.3pp. 1 of 7 models improved week over week; issue #4's leader deepseek-v4-pro is now 2 of 7.
- **No ticker cleared 51%.** BNB is the week's most readable coin (47.3%, n=916) and SOL the hardest at 38.7%; 0 of the 5 assets cleared 51%. Pair of the week: deepseek-v4-pro x BNB 53.5% (n=144); worst: grok-4.6 x SOL 29.5% (n=105).
- **0 of 2 qualifying turns found a caller.** ETH long +4.61%, 0 caller(s), first NOBODY; ETH short -2.13%, 0 caller(s), first NOBODY.
- **Self-agreement read as a penalty in every cell.** Cross-TF/FH lifts ran -17.4pp to -2.5pp (issue #4: -7.6pp to -1.9pp). On trend days, 4h TF-agreement hit 33.6% (n=214) against 36.8% (n=247) 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 42.9% -> 42.8%. The grok column compares this window's pure 4.6 line against issue #4's spliced 4.5+4.6 line. "Was" values are the numbers published in issue #4; 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 #5's grok row is a pure grok-4.6 week while issue #4's was a lineage splice, so read the grok week-over-week 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 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits inside the PREVIOUS window (week A, issue #4): every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)"; this window is pure grok-4.6, so the grok week-over-week row compares a pure-4.6 week with a spliced week.

Leaderboard of the week

Model	N	Cov.	Hit rate	95% CI	W/w pp
gemini-3.1-pro	761	57.3%	46.5%	43.0%-50.1%	+2.3
deepseek-v4-pro	724	54.6%	44.2%	40.6%-47.8%	-0.4
claude-fable-5	628	47.2%	42.8%	39.0%-46.7%	-0.2
qwen-3.8-max	697	53.5%	42.5%	38.9%-46.2%	-0.1
claude-opus-5	545	41.1%	42.4%	38.3%-46.6%	+0.0
gpt-5.6-sol	689	51.8%	41.6%	38.0%-45.4%	-1.0
grok-4.6	476	36.0%	37.6%	33.4%-42.0%	-2.3

No weekly title is awarded this issue -- for the fifth time. Title rule: ≥ 5 pp gap AND non-overlapping 95% CI, $N \geq 10$ per side. gemini-3.1-pro's lead over deepseek-v4-pro is 2.3pp and their CIs overlap, so neither condition is met. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits inside the PREVIOUS window (week A, issue #4): every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)"; this window is pure grok-4.6, so the grok week-over-week row compares a pure-4.6 week with a spliced week.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #4
BNB	916	47.3%	44.1%-50.5%	41.2%
XRP	909	44.2%	41.0%-47.5%	40.3%
ETH	982	43.1%	40.0%-46.2%	43.3%
BTC	780	40.6%	37.2%-44.1%	43.7%
SOL	933	38.7%	35.6%-41.9%	45.8%

BNB was the field's most readable ticker (47.3%) and SOL the hardest (38.7%); 0 of 5 assets cleared 51% this week (issue #4: 0 of 5). Issue #4's best ticker was SOL and its hardest XRP -- this issue's ordering is BNB > XRP > ETH > BTC > SOL.

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
deepseek-v4-pro	BNB	144	53.5%
claude-opus-5	BNB	112	49.1%
claude-fable-5	BNB	127	48.0%
gemini-3.1-pro	BNB	155	47.7%
deepseek-v4-pro	ETH	164	46.9%
gemini-3.1-pro	XRP	147	46.9%

Model (5 worst)	Symbol	N	Hit rate
claude-opus-5	SOL	112	36.6%
deepseek-v4-pro	SOL	148	36.5%
grok-4.6	BTC	70	35.7%
grok-4.6	ETH	99	35.4%
grok-4.6	SOL	105	29.5%

deepseek-v4-pro x BNB (53.5%, n=144) is the week's best pair and grok-4.6 x SOL (29.5%, n=105) the weakest. Pair history accumulates across issues -- read this as the fifth 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
ETH	Thu Sep 3, 08:00	Long	+4.61%	0	NOBODY
ETH	Fri Sep 4, 08:00	Short	-2.13%	0	NOBODY

The ETH long turn drew NOBODY; the ETH short turn drew NOBODY. Scope in v1 remains BTC/ETH only; the event count across five 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	461	35.4%	10	90.0%	39.9% (651)	-4.6
cross-TF	1d	TF 1d vs 4h	67	28.4%	12	83.3%	43.8% (128)	-15.4
cross-FH	4h	by FH 1h	401	37.4%	20	75.0%	39.9% (651)	-2.5
cross-FH	1d	by FH 4h	57	26.3%	12	83.3%	43.8% (128)	-17.4

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 177 / 49 / 227 / 59. Pooled across all four cell families the lifts run -17.4pp to -2.5pp this week (issue #4: -7.6pp to -1.9pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has 3 trend days of 7: 08-31 +1.19%, 09-01 -1.50% (trend), 09-02 -0.06%, 09-03 +4.98% (trend), 09-04 -1.97% (trend), 09-05 +0.24%, 09-06 +0.58%. On trend days, 4h TF-agreement hit 33.6% (n=214) against 36.8% (n=247) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	32.6%	14.6%	52.8%	2.24
claude-opus-5	1,326	29.5%	11.6%	58.9%	2.54
deepseek-v4-pro	1,325	31.2%	23.4%	45.4%	1.34
gemini-3.1-pro	1,329	31.6%	25.7%	42.7%	1.23
gpt-5.6-sol	1,330	26.4%	25.4%	48.2%	1.04
grok-4.6	1,322	15.4%	20.6%	64.0%	0.74
qwen-3.8-max	1,302	34.0%	19.5%	46.5%	1.74

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 1.55 this week. 'Sideways' was 51.2% of mature forecasts (4,744 of 9,264). Consensus skew: 521 symbol/FH/TF cells had 5 or more models on the same side and hit 41.3% (n=3,248) against the 42.8% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	BTC	Long	80	4h / 4h	Sep 4, 12:01	SL	-\$1.60

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

Calibration bridge

The hit-rate leader this week is gemini-3.1-pro (46.5%); the best-calibrated line is qwen-3.8-max (Brier 0.2755, gap +16.6pp). Per-model ok-rate this window ran 98.10% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #5.

Market check: an up week on lower volume

OBSERVATION — mean |1d move| 2.35% -> 2.10% (-11% rel), BTC realized vol 37.9% -> 35.2% (ann., hourly); field directional accuracy 42.9% -> 42.8% (-0.1pp), 1 of 7 models improved, sim win-rate up for 1 of 7, field sim PnL -\$473.59 -> -\$669.03 (adjacent calendar weeks Aug 24-30 vs Aug 31-Sep 6; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 44.2% -> 51.6% — the opposite direction to the trade-based hit rule.

Measure	Aug 24-30 (week A)	Aug 31-Sep 6 (week B)	Change
Mean 1d move (5 assets)	2.35%	2.10%	-11% rel
BTC realized vol (ann., hourly)	37.9%	35.2%	-2.7pp
Field directional accuracy	42.9%	42.8%	-0.1pp
Models improving hit-rate	--	1 of 7	--
Raw price-sign accuracy	44.2%	51.6%	+7.4pp
Field sim win-rate	37.0%	35.6%	-1.4pp
Field sim net PnL	-\$473.59	-\$669.03	--

Week A is exactly the issue-#4 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 A). Week A (2026-08-24..2026-08-30) mixes the 4.5-era (Aug 24 00:00-09:22) with 4.6 and reproduces the published issue #4; week B is pure grok-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. Five issues, and the best ticker has changed hands every time (SOL -> BNB -> BTC -> SOL -> BNB).
- Self-agreement lifts run -17.4pp to -2.5pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Aug 31-Sep 6, 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 #4: 3 of 7); every comparison with issue #4 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. 10 events across five issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=10, 12, 20, 12); 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, qwen-3.8-max. Series density and lineage notes (wave 5): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 31-Sep 6) daily coverage is FULL, as it was in issue #4 (the first full week of 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 490, 1M 490). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, inside the PREVIOUS window (week A, issue #4): this window carries 1,470 grok rows and none from the 4.5 era, so every grok number in this issue is a pure grok-4.6 line, while every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)" -- the grok week-over-week row compares a pure-4.6 week with a spliced week. (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 #4 after the audit that found the raw table mixes two exchanges; issue #4 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 (Sep 7 16:00 UTC) -- the definition pinned in issue #3 and carried by issue #4, which printed 38,385 at the Aug 31 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 42,905 is an additive step under one definition; counter deltas against issue #3 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,264
OK in gate	9,264	-- of them directional	4,520
Out of gate (1w / 1M)	490 / 490	-- of them sideways	4,744
Invalid	46	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-08-31.json	Report cutoff	Mon Sep 7, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-08 08:08 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,520 scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 7, 2026 cutoff): 42,905 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 25 published reports.

Issue #5. Weekly Model Watch is a living series. Issue #4's open questions -- does any model hold the top rank two issues running and does the title rule ever fire, and does the reversal detector find callers outside ETH? -- read this issue as: gemini-3.1-pro at the top (46.5%) where issue #4 had deepseek-v4-pro, no weekly title was awarded, for the fifth issue running, and 0 of 2 qualifying turns found a caller, on turns in ETH only. The grok row is pure grok-4.6 this issue; issue #4's grok row was the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)". 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 at all, and on a non-ETH turn?
- 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 #5	https://marketmania.ai/research/reports/consensus-watch-2026-08-31.pdf
Weekly Calibration #5	https://marketmania.ai/research/reports/weekly-calibration-2026-08-31.pdf
Weekly Model Watch #4	https://marketmania.ai/research/reports/model-watch-2026-08-24.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_2026w36,
  title = {Weekly Model Watch #5: weekly leaderboard, ticker/pair reads and cross-confirmation, Aug 31-Sep 6 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {8},
  url = {https://marketmania.ai/research/reports/model-watch-2026-08-31.pdf},
  note = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Aug 31-Sep 6, 2026 UTC; source weekly_metrics_2026-08-31.json}
}
```