

Weekly Model Watch #2

August 19, 2026 · window Aug 10-16 · MarketMania Research · Weekly series

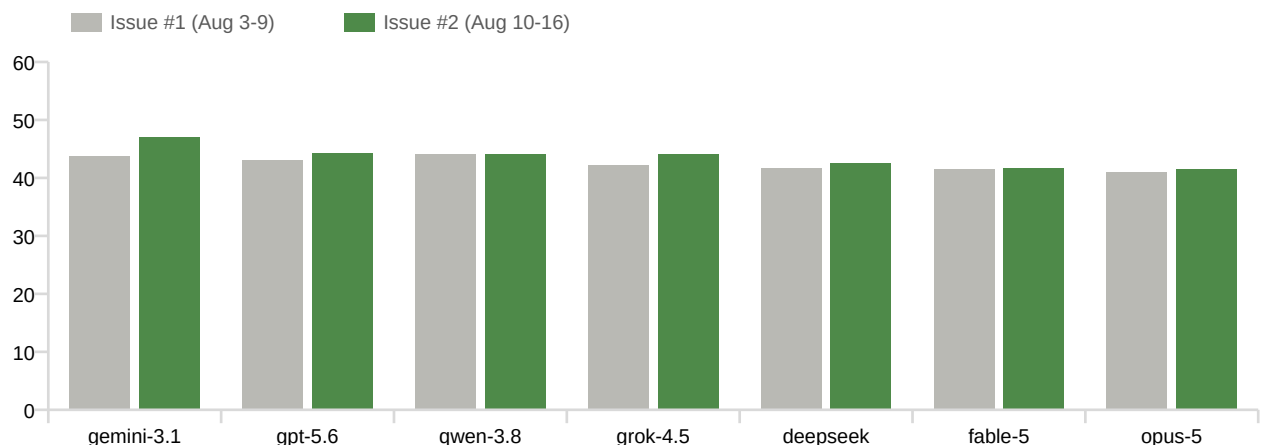
RESEARCH SNAPSHOT			
window Aug 10-16, 2026 UTC	7 models (stable lineup)	3,670 directional calls scored of 9,307 mature	FH gate 1h / 4h / 1d
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	base field hit 43.8%	cutoff Mon Aug 17, 2026, 16:00 UTC	methodology v1.1 (2026-08-10) , hash e66c7e8c864a2233
market: BTC net -3.08% (prior +2.09%)	ann. vol 10.0% (was 11.4%)	TOP5 volume \$7.75B , -8.9% w/w	pairwise corr 0.52 (was 0.40)

KEY FINDING · OBSERVATION (one weekly window)

The reversal blind spot repeated: zero of 7 models caught ETH's -2.65% Monday turn -- two issues, two qualifying reversals, zero callers.

TL;DR

- **OBSERVATION -- No weekly title again.** gemini-3.1-pro leads descriptively at **47.0%** [42.9%, 51.2%] after the field's biggest jump (+3.3pp w/w), but the gap to #2 (gpt-5.6-sol, 44.2%) is 2.8pp with overlapping CIs -- the title rule (≥ 5 pp gap AND non-overlapping 95% CI) was not met. Six of seven models improved week-over-week; last week's leader qwen-3.8-max held flat at 44.1% and slipped 1st -> 3rd.
- **Ticker readability reshuffled hard.** BNB is the week's most readable coin (49.3%, n=816) and last week's best, SOL, collapsed 48.2% -> 38.9% -- a -9.3pp swing to 4th of 5. ETH is the hardest for the second week running (35.6%). Pair of the week: gemini-3.1-pro x BNB 55.7% (n=122); worst: deepseek-v4-pro x ETH 29.9% (n=87).
- **The reversal blind spot repeated.** The week's only qualifying reversal -- ETH turning short at 08:00 UTC on Monday Aug 10, an 8-hour move of -2.65% -- drew zero callers from all 7 models. Two issues, two reversals (both on the window's Monday), first caller: NOBODY, twice.
- **Self-agreement still added nothing pooled -- but the first regime split landed.** Cross-TF/FH lifts ran -18.5 to +0.2pp, and 1d confirmation was a disaster (17.9-18.5% hit). On the window's single trend day, though, 4h TF-agreement hit 54.3% vs 38.3% on the six flat days (n=46 vs 264) -- the regime control's first non-degenerate read; a hint, not a result.



Directional hit-rate (%) by model, week over week, ranked by this week's result. Field base this week: 43.8% (issue #1: 42.4%). Six of seven models improved; qwen-3.8-max held flat (44.1% both weeks) and gemini-3.1-pro's +3.3pp is the field's largest move. The legacy qwen-3.7-max row retired with issue #1, so both weeks plot the same 7 models.

Why it matters

MarketMania scores 7 models 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. With issue #2 the series gains its first week-over-week movement column -- and its first non-degenerate regime control.

Leaderboard of the week

The full 7-model field, ranked by this week's directional hit-rate; W/w pp = movement vs issue #1 (begins this issue, as promised).

Model	N	Coverage	Hit rate	95% CI	W/w pp
gemini-3.1-pro	555	41.7%	47.0%	42.9%-51.2%	+3.3
gpt-5.6-sol	636	47.8%	44.2%	40.4%-48.1%	+1.2
qwen-3.8-max	572	43.0%	44.1%	40.0%-48.1%	0.0
grok-4.5	620	46.6%	44.0%	40.2%-48.0%	+1.8
deepseek-v4-pro	454	34.2%	42.5%	38.0%-47.1%	+0.9
claude-fable-5	472	35.5%	41.7%	37.4%-46.2%	+0.3
claude-opus-5	361	27.1%	41.5%	36.6%-46.7%	+0.6

No weekly title is awarded this issue -- for the second time. Title rule: **>=5pp gap AND non-overlapping 95% CI, N>=10.**

gemini-3.1-pro's lead over gpt-5.6-sol is 2.8pp (47.0% vs 44.2%) and their CIs overlap, so neither condition is met -- the lead is descriptive. Rank shuffle vs issue #1: gemini-3.1-pro 2nd -> 1st, qwen-3.8-max 1st -> 3rd (flat hit-rate in a rising field), grok-4.5 4th -> 4th. Hit-rate rank and Brier (calibration) rank still diverge: the hit-rate leader gemini-3.1-pro is only 5th of 7 by Brier -- see the calibration bridge below and Weekly Calibration #2.

Ticker of the week

This week's 5-ticker field, ranked by directional hit-rate.

Symbol	N	Hit rate	95% CI
BNB	816	49.3%	45.9%-52.7%
XRP	824	48.5%	45.1%-52.0%
BTC	727	44.0%	40.5%-47.6%
SOL	659	38.9%	35.2%-42.6%
ETH	644	35.6%	32.0%-39.3%

BNB was the field's most readable ticker this week (49.3%, n=816) and ETH the hardest for the second issue running (35.6%, n=644); their 95% CIs are well separated (45.9-52.7% vs 32.0-39.3%), while every adjacent pair in the ranked list overlaps its neighbor. The reshuffle is the story: last week's best, SOL (48.2%), fell to 4th at 38.9% -- a -9.3pp swing -- and last week's 3rd, BNB, took the top. Read per-ticker readability as weekly weather, not a durable per-asset skill claim; ETH-hardest is the only repeat so far.

Pair of the week

Top 6 of the 12 model x ticker pairs tracked this week (best hit-rate), and the 5 worst. Pairs shown require N>=8 calls.

Top 6 pairs

Model	Symbol	N	Hit rate
gemini-3.1-pro	BNB	122	55.7%
qwen-3.8-max	BNB	124	50.8%
gemini-3.1-pro	XRP	133	50.4%
gpt-5.6-sol	BNB	145	49.7%
qwen-3.8-max	XRP	127	49.6%
deepseek-v4-pro	XRP	97	49.5%

5 worst pairs

Model	Symbol	N	Hit rate
deepseek-v4-pro	SOL	78	35.9%
gpt-5.6-sol	ETH	113	35.4%
claude-opus-5	ETH	65	35.4%
claude-fable-5	ETH	84	33.3%
deepseek-v4-pro	ETH	87	29.9%

gemini-3.1-pro x BNB (55.7%, n=122) is the week's best pair -- the only pairing above 55% -- and the same model tops the leaderboard. deepseek-v4-pro sits at both ends: x XRP makes the top 6 (49.5%) while x SOL and x ETH (29.9%, the week's worst) anchor the bottom. Four of the five worst pairs are ETH pairings. Pair history accumulates across issues -- read this as the second data point, not a ranking.

Who saw the reversal first

Reversal definition: 4h grid; trend = sign of prior 24h; counter-move $\geq 2.0\%$ (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	Mon Aug 10, 08:00	Short	-2.65%	0	NOBODY

This week's only qualifying reversal was ETH's turn to short at 08:00 UTC on Monday, Aug 10 -- an 8-hour move of -2.65%. Zero of the 7 active models had a matured hit call in the new (short) direction inside the window; first caller: **NOBODY** -- the same outcome as issue #1's BTC turn. Two issues, two reversals, zero callers; both events landed on the window's Monday. Still a two-event sample: reported descriptively, not as a claim about turn-detection skill. Scope in v1 remains BTC/ETH only.

Cross-TF confirmation: does agreeing with yourself help?

Within one forecast horizon (FH), does a model's call from a longer input timeframe (TF) agree with its call from a shorter one -- and if so, does the call do any better?

FH	vs.	Agree n	Agree hit	Dis n	Dis hit	Mixed	Base hit	Lift pp
4h	TF 4h vs 1h	310	40.6% [35.3%-46.2%]	20	35.0% [18.1%-56.7%]	247	41.2% (n=577)	-0.6
1d	TF 1d vs 4h	56	17.9% [10.0%-29.8%]	0	n/a (0)	43	36.4% (n=99)	-18.5

n in Base hit column is the base-rate call count. * = N<10 (insufficient); # = N<20 (thin); "n/a (0)" = no disagreeing calls. Insufficient/thin cells are never used to rank models. The 1d row is the week's ugliest number: when a model's 1d call was confirmed by its own 4h-TF read, it hit 17.9% -- 18.5pp BELOW the 1d base.

Regime control (first non-degenerate split). Methodology calls for a trend-vs-flat split on this table. Issue #1 had zero trend days; this window has exactly one -- Mon Aug 10 (-1.63%), the other six days were flat. On that single trend day, 4h TF-agreement hit **54.3%** (n=46) vs **38.3%** on flat days (n=264); the disagree cells ran 16.7% (n=6) vs 42.9% (n=14). One trend day is a hint that self-agreement may be regime-dependent -- exactly what the monthly block-bootstrap test is for -- not a result to trade.

Cross-FH confirmation: does a shorter horizon confirm a longer one?

Does a model's call at one forecast horizon (FH) agree with its call at a shorter FH on the same timeframe -- and if so, does the call do any better?

FH	vs.	Agree n	Agree hit	Dis n	Dis hit	Mixed	Base hit	Lift pp
4h	by FH 1h	299	41.5% [36.0%-47.1%]	31	29.0% [16.1%-46.6%]	246	41.2% (n=577)	+0.2
1d	by FH 4h	54	18.5% [10.4%-30.8%]	2	100.0% [34.2%-100.0%] *	43	36.4% (n=99)	-17.8

The 4h FH-confirmation is the only cell family near breakeven this week (+0.2pp pooled; 48.9% on the trend day vs 40.2% flat). The 1d rows repeat the cross-TF story: 18.5% agree-hit, -17.8pp lift. Disagree cells stay tiny (n=0, 2, 20, 31); the 1d disagree pair (n=2, both hits) is insufficient and never used to rank models.

Side-mix and the week's failure

How each model split its calls this week across long / short / sideways, and the week's single costliest miss.

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	19.6%	15.9%	64.5%	1.23
claude-opus-5	1,330	14.5%	12.6%	72.9%	1.15
deepseek-v4-pro	1,328	10.5%	23.7%	65.8%	0.44
gemini-3.1-pro	1,330	16.2%	25.5%	58.3%	0.64
gpt-5.6-sol	1,330	14.7%	33.1%	52.2%	0.45
grok-4.5	1,330	15.9%	30.8%	53.4%	0.52
qwen-3.8-max	1,329	17.4%	25.7%	57.0%	0.68

L:S = each model's own long-share divided by its short-share. The field flipped its skew: issue #1 ran ~2.2:1 long-over-short; this week the average of the 7 per-model ratios is ~0.7:1 -- short-over-long -- in a week BTC fell -3.08%. gpt-5.6-sol is the most short-committed (33.1% short vs 14.7% long); claude-fable-5 and claude-opus-5 are the only models still leaning long. 'Sideways' stayed the field's most common call: 60.6% of mature forecasts (5,637 of 9,307), up from 55.7%.

Consensus skew. 404 symbol/FH/TF cells this week had 5 or more of the (up to 7) active models on the same side. Those skewed cells' mature calls hit 41.8% (n=2,486) -- below the week's 43.8% field base rate, so leaning with a 5-of-7+ skew did not outperform the field for the second week running. (Strict unanimity is Consensus Watch's own metric -- see that report for the full herding analysis.)

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gpt-5.6-sol	ETH	Short	82	4h / 1h	Aug 10, 16:01	Expiry	-\$0.36

Fail of the week. The week's single highest-confidence individual miss: gpt-5.6-sol, ETH, short, confidence 82, FH 4h / TF 1h, slot Aug 10 16:01 UTC, exit by expiry, net -\$0.36. Cross-reference: ETH's only qualifying reversal (Block 4) had turned short at 08:00 that same day -- this conf-82 short entered at 16:01, after the qualifying 8-hour counter-move had already completed, and expired without reaching TP. ETH was also the week's hardest ticker (35.6%), and 4 of the 5 worst pairs are ETH pairings.

Counters and the calibration bridge

Counter	Value	Counter	Value
Forecasts total	9,380	Mature (scored pool)	9,307
OK in gate	9,307	-- of them directional	3,670
Out of gate (1w)	69	-- sideways	5,637
Invalid	4	Pending (next issue)	0
Uptime (slots seen/expected)	168/168 · 42/42 · 42/42 · 7/7 · 7/7	Late closes	0
Source file	weekly_metrics_2026-08-10.json	Report cutoff	Mon Aug 17, 16:00 UTC
Generated at (pipeline)	2026-08-18 08:37 UTC	Methodology	v1.1 · hash e66c7e8c864a2233

Data quality. First perfect-uptime week of the series: every 1h, 4h and 1d slot grid ran full (issue #1: 165/168 on the hourly grid). Per-model ok-rate was 99.85% for qwen-3.8-max and deepseek-v4-pro and 100.0% for the other five.

Calibration bridge. qwen-3.8-max is again the week's best-calibrated model by Brier score (0.2688, lowest of the 7) -- but this issue it is only 3rd by hit-rate, while the hit-rate leader gemini-3.1-pro ranks 5th of 7 by Brier (0.2848) with a +18.4pp overconfidence gap. Hit-rate and calibration keep rewarding different behavior; the full confidence-bucket analysis (including the field's first working 70-80 bucket) lives in Weekly Calibration #2.

Practical implications

- Do not read this week's leaderboard as a skill ranking: no title was awarded, the top CIs overlap, and gemini-3.1-pro's +3.3pp jump is one week of movement in a field where last week's leader just went flat.
- Treat per-ticker readability as weekly weather. The best coin flipped (SOL -> BNB, a -9.3pp swing for SOL); the only repeat so far is ETH-hardest, two weeks running -- and the field missed both Monday reversals outright, which still points at turns as the general weak spot.
- Do not use within-model cross-TF/FH agreement as a confidence booster: pooled lifts ran -18.5 to +0.2pp, and 1d confirmation was the week's worst cell family (~18% hit). The one trend day's 54.3% agree-hit is the thing to watch at monthly n, not a signal to act on.

Limitations

- Single week (Aug 10-16, 2026 UTC). Every finding is descriptive for this window only; no claim is made about next week or any model's underlying skill.
- The regime control rests on ONE trend day (Mon Aug 10, n=46/45 agree calls); the other six days were flat. The split is reported because it is the series' first non-degenerate one, not because one day can establish anything.
- Observations inside one window are not independent: the same models watch overlapping symbol/FH/TF cells hour after hour; agree/disagree calls made close in time are correlated, not i.i.d. draws.
- 95% Wilson CIs shown throughout are descriptive, not inferential, for the reason above -- read them as a range, not a formal coverage guarantee.
- The reversal detector remains v1 and BTC/ETH-only, on price series reconstructed from trade entry prices (median per symbol-slot), not an independent tick feed. Two events in two issues cannot characterize detector performance either way.
- Cross-TF and cross-FH disagreement cells are tiny again (n=0, 2, 20, 31) -- cells below N=10 are marked insufficient and are never used to rank models.
- Week-over-week movement begins this issue on two data points; rank shuffles at these CI widths are expected noise until several issues accumulate.

THIS REPORT: 3,670 scored observations. **MARKETMANIA RESEARCH TO DATE** (as of Aug 17, 2026 cutoff): 23,458 resolved forecasts since Jul 11 · 9 models tracked (7 current frontier + 2 archived legacy) · 5 assets · 5 forecast horizons · hourly cadence · 9 published research reports incl. this wave.

Issue #2. Weekly Model Watch is a living series; the movement column and the regime control went live this issue. Issue #1's open questions -- does the reversal blind spot repeat, and does cross-TF agreement stay flat-to-negative? -- both closed 'yes': a second Monday reversal drew zero callers, and pooled self-agreement lifts stayed at -18.5 to +0.2pp (with the first trend-day hint of regime dependence). Engine 1.1 powers the sandbox since Aug 18 (after this window closed); no cross-engine PnL comparisons are claimed.

What we're testing next

- **Next issue:** a third Monday reversal would make the blind spot a pattern; does SOL's readability bounce back or was issue #1 the outlier; does the trend-day agreement premium reappear?
- **Monthly test (Sep 2):** cross-TF/FH lifts with regime control across mixed weeks (trend vs. flat), block bootstrap.

Related research

Consensus Watch #2 -- <https://marketmania.ai/research/reports/consensus-watch-2026-08-10.pdf>

Weekly Calibration #2 -- <https://marketmania.ai/research/reports/weekly-calibration-2026-08-10.pdf>

Config Watch #2 -- <https://marketmania.ai/research/reports/config-watch-2026-08-12.pdf>

Weekly Model Watch #1 -- <https://marketmania.ai/research/reports/model-watch-2026-08-03.pdf>

The three weekly reports publish together as one issue each week; Config Watch follows on its own cycle.

CITE THIS REPORT

```
@misc{mm_model_watch_2026w33,
  title = {Weekly Model Watch #2: weekly leaderboard, ticker/pair reads and cross-confirmation, Aug 10-16 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {August}, day = {19},
  url = {https://marketmania.ai/research/reports/model-watch-2026-08-10.pdf},
  note = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Aug 10-16, 2026 UTC; source weekly_metrics_2026-08-10.json}
}
```