

Weekly Model Watch #3

August 26, 2026 · MarketMania Research · Weekly series · window Aug 17-23, 2026 UTC

RESEARCH SNAPSHOT

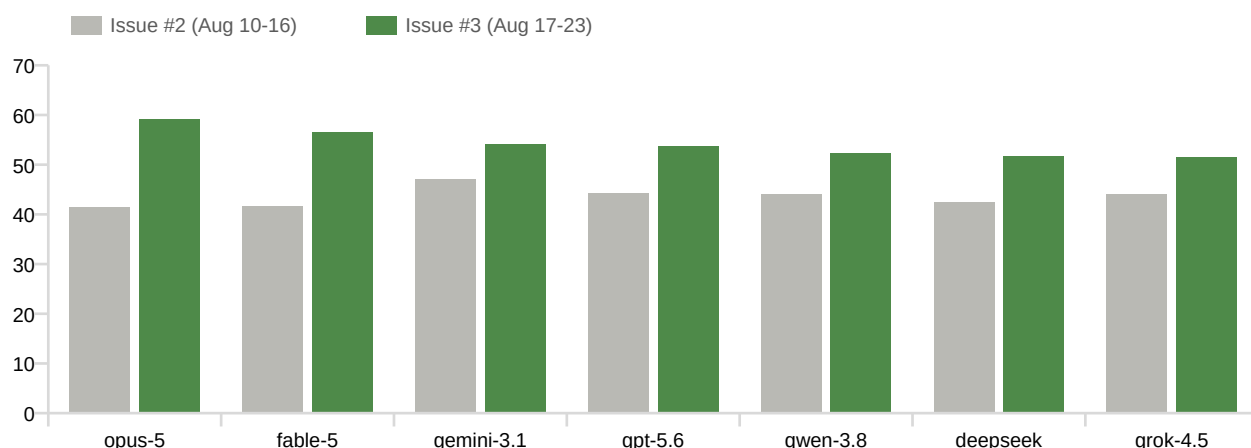
5,130 directional calls scored	7 models (stable lineup)	of 9,260 mature forecasts	window Aug 17-23, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 54.1%	cutoff Mon Aug 24, 2026, 16:00 UTC
market: BTC net +23.58% (prior -3.08%)	ann vol 71.1% (was 10.0%)	TOP5 volume \$25.97B , +235% w/w	pairwise corr 0.81 (was 0.52)

KEY FINDING · OBSERVATION (one weekly window)

The market moved and the models kept up: field accuracy 43.8% -> 54.1% (+10.3pp) with 7 of 7 models improving -- and after two blind weeks, both ETH reversals finally found callers.

TL;DR

- **OBSERVATION -- No weekly title again.** claude-opus-5 leads descriptively at 59.1% [55.3%, 62.7%] after the field's biggest jump (+17.5pp w/w), but the gap to #2 (claude-fable-5, 56.5%) is 2.6pp with overlapping CIs -- the title rule (≥ 5 pp gap AND non-overlapping 95% CI) was not met for the third issue running. All 7 models improved; issue #2's leader gemini-3.1-pro slipped 1st -> 3rd on a +7.1pp gain.
- **Every ticker was readable.** BTC is the week's most readable coin (57.5%, n=1,118) and XRP the hardest at 51.2% -- the first week in the series where every one of the 5 assets cleared 51%. Issue #2's hardest ticker, ETH, rose 35.6% -> 52.8%; SOL bounced 38.9% -> 54.3%. Pair of the week: claude-opus-5 x BNB 60.3% (n=126); worst: grok-4.5 x XRP 45.2% (n=135).
- **The reversal blind spot broke.** Both qualifying reversals were ETH: the Aug 22 turn short (-3.46% over 8h) drew 1 caller, first gemini-3.1-pro; the Aug 23 turn long (+2.14%) drew 3, first claude-fable-5. Issues #1 and #2 had two reversals and zero callers between them.
- **Self-agreement turned mildly positive -- and the regime split widened.** Cross-TF/FH lifts ran -0.6 to +4.6pp (issue #2: -18.5 to +0.2pp), with the 1d cross-TF cell going from the week's disaster (17.9%) to its best agree bucket (75.9%, n=133). On the 5 trend days, 4h TF-agreement hit 64.8% vs 47.5% on the 2 flat days (n=455 vs 120).



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 43.8% -> 54.1%. "Was" values are the numbers published in issue #2; 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 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. Issue #3 is the series' first live-tape week -- 5 trend days against issue #2's 1 -- so every week-over-week movement here is a movement across regimes as well as across weeks.

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.

Leaderboard of the week

Model	N	Cov.	Hit rate	95% CI	W/w pp
claude-opus-5	672	50.5%	59.1%	55.3%-62.7%	+17.5
claude-fable-5	779	58.6%	56.5%	53.0%-59.9%	+14.7
gemini-3.1-pro	798	60.1%	54.1%	50.7%-57.6%	+7.1
gpt-5.6-sol	688	51.7%	53.6%	49.9%-57.3%	+9.4
qwen-3.8-max	706	55.1%	52.3%	48.6%-55.9%	+8.2
deepseek-v4-pro	777	58.4%	51.6%	48.1%-55.1%	+9.1
grok-4.5	710	53.4%	51.4%	47.7%-55.1%	+7.4

No weekly title is awarded this issue -- for the third time. Title rule: ≥ 5 pp gap AND non-overlapping 95% CI, $N \geq 10$ per side. claude-opus-5's lead over claude-fable-5 is 2.6pp and their CIs overlap, so neither condition is met. Rank shuffle vs issue #2: claude-opus-5 7th -> 1st, claude-fable-5 6th -> 2nd, gemini-3.1-pro 1st -> 3rd. Both Claude models made the field's two biggest jumps from the bottom two ranks, in a week when the base rate moved +10.3pp under everyone.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #2
BTC	1,118	57.5%	54.6%-60.4%	44.0%
SOL	1,036	54.3%	51.3%-57.4%	38.9%
BNB	932	54.0%	50.8%-57.2%	49.3%
ETH	1,074	52.8%	49.8%-55.8%	35.6%
XRP	970	51.2%	48.1%-54.4%	48.5%

BTC was the field's most readable ticker (57.5%) and XRP the hardest (51.2%); their CIs are separated, while every adjacent pair in the middle of the list overlaps its neighbour. The story is the floor, not the top: this is the first week where all 5 assets cleared 51%, and both previous extremes reversed -- issue #2's best (BNB) is 3rd and issue #2's worst (ETH, hardest for two issues running) is 4th. Three issues in, no asset has held a rank.

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
claude-opus-5	BNB	126	60.3%
grok-4.5	BTC	141	60.3%
gemini-3.1-pro	BTC	175	59.4%
claude-opus-5	ETH	138	59.4%
claude-opus-5	SOL	137	59.1%
claude-opus-5	XRP	119	58.8%

Model (5 worst)	Symbol	N	Hit rate
qwen-3.8-max	ETH	140	49.3%
grok-4.5	ETH	159	49.1%
deepseek-v4-pro	ETH	168	48.2%
qwen-3.8-max	XRP	144	47.9%
grok-4.5	XRP	135	45.2%

claude-opus-5 x BNB (60.3%, n=126) is the week's best pair -- by 0.04pp over grok-4.5 x BTC, which rounds to the same 60.3% -- and claude-opus-5 takes 4 of the top 6. At the other end every one of the 5 worst pairs still cleared 45%, and 3 of the 5 are ETH pairings (issue #2: 4 of 5). Pair history accumulates across issues -- read this as the third 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	Sat Aug 22, 00:00	Short	-3.46%	1	gemini-3.1-pro
ETH	Sun Aug 23, 08:00	Long	+2.14%	3	claude-fable-5

Two qualifying reversals, both ETH, and both found callers -- the first non-zero rows in the series. The Aug 22 turn drew exactly one caller: gemini-3.1-pro, on a 4h/1h call placed at 00:01 with confidence 60. The Aug 23 turn drew three -- claude-fable-5 (first, 1d/1d slot at 00:01, confidence 62), claude-opus-5 and deepseek-v4-pro. Two weeks of NOBODY became one week of 1 and 3 callers; that is four qualifying events across three issues, still far too few to characterise turn-detection skill either way. Scope in v1 remains BTC/ETH only.

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	575	61.2%	41	41.5%	60.6% (840)	+0.6
cross-TF	1d	TF 1d vs 4h	133	75.9%	5 *	40.0%	71.4% (171)	+4.6
cross-FH	4h	by FH 1h	508	60.0%	42	47.6%	60.6% (840)	-0.6
cross-FH	1d	by FH 4h	100	74.0%	11	72.7%	71.4% (171)	+2.7

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 223 / 33 / 288 / 60. The 1d cross-TF row is the mirror image of issue #2 -- a 1d call confirmed by the model's own 4h-TF read hit 75.9%, 4.6pp above the 1d base, where the same cell hit 17.9% and -18.5pp last week. Pooled across all four cell families the lifts now run -0.6 to +4.6pp: no longer a penalty, not yet an edge.

Regime control. Methodology calls for a trend-vs-flat split on this table. Issue #1 had zero trend days, issue #2 exactly one; this window has five: 08-17 +2.55% (trend), 08-18 +0.21%, 08-19 +7.82% (trend), 08-20 +5.07% (trend), 08-21 +7.26% (trend), 08-22 -1.67% (trend), 08-23 +0.69%. On the 5 trend days, 4h TF-agreement hit 64.8% (n=455) vs 47.5% on the 2 flat days (n=120); the disagree cells ran 41.0% (n=39) vs 50.0% (n=2, insufficient). The 1d cell splits the other way (74.3% trend, n=109, vs 83.3% flat, n=24). Issue #2's single-trend-day hint (54.3% vs 38.3%) reappears on the 4h cell at a readable n -- the first time the regime control has had two

usable sides. Still one window; the monthly block-bootstrap test is where this belongs.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	52.3%	6.3%	41.4%	8.27
claude-opus-5	1,330	45.9%	4.6%	49.5%	10.01
deepseek-v4-pro	1,330	46.6%	11.8%	41.6%	3.95
gemini-3.1-pro	1,329	45.8%	14.3%	40.0%	3.20
gpt-5.6-sol	1,330	38.4%	13.4%	48.3%	2.87
grok-4.5	1,330	35.2%	18.2%	46.6%	1.93
qwen-3.8-max	1,281	45.4%	9.7%	44.9%	4.69

L:S = each model's own long-share divided by its short-share. The field flipped its skew back and then some: issue #1 ran ~2.2:1 long-over-short, issue #2 ~0.7:1 short-over-long, and this week the average of the 7 per-model ratios is ~5.0:1 long-over-short in a week of 5 trend days. claude-fable-5 carried the highest long share (52.3%) and claude-opus-5 the highest ratio (10.01); grok-4.5 is the most balanced (1.93) and the field's weakest hit-rate. 'Sideways' fell to 44.6% of mature forecasts (4,130 of 9,260) from 60.6%, which is why the directional pool grew from 3,670 to 5,130. Consensus skew: 593 symbol/FH/TF cells had 5 or more models on the same side and hit 54.4% (n=3,683) -- just above the 54.1% base, the first issue where a 5-of-7+ skew did not underperform the field, though 0.4pp is well inside noise.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	XRP	Long	85	4h / 1h	Aug 21, 16:01	SL	-\$3.10

Fail of the week. The week's single highest-confidence individual miss. Cross-reference: XRP was the week's hardest ticker (51.2%) and the asset with the largest mean |1d| move (6.65%, against BTC's 3.34%) -- an 85-confidence directional call into the week's noisiest tape. Unlike issue #2's expiry-based fail, this one hit its stop, which is what the calibration v2 TP/SL work is aimed at.

Calibration bridge

For the first time in the series the hit-rate leader and the best-calibrated model are the same: claude-opus-5 tops the leaderboard at 59.1% and posts the field's lowest Brier (0.2419) with a +2.2pp overconfidence gap. Issue #2's best-calibrated model, qwen-3.8-max, is 3rd by Brier (0.2540) and 5th by hit-rate. Per-model ok-rate this window was 100.0% for five models, 99.93% for gemini-3.1-pro and 96.20% for qwen-3.8-max. The full confidence-bucket analysis lives in Weekly Calibration #3.

Market check: the tape woke up

OBSERVATION — the tape sped up and accuracy followed: mean |1d move| 0.73% -> 4.28% (483% rel), BTC realized vol 19.3% -> 58.7% (ann., hourly); field directional accuracy 43.8% -> 54.1% (+10.3pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -\$527 -> +\$889 (adjacent calendar weeks Aug 10-16 vs Aug 17-23; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 40.1% -> 53.5% — consistent with the trade-based hit rule.

Measure	Aug 10-16 (week A)	Aug 17-23 (week B)	Change
Mean 1d move (5 assets)	0.73%	4.28%	+483% rel
BTC realized vol (ann., hourly)	19.3%	58.7%	+39.4pp
Field directional accuracy	43.8%	54.1%	+10.3pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	40.1%	53.5%	+13.4pp
Field sim win-rate	29.7%	46.3%	+16.6pp
Field sim net PnL	-\$527.07	+\$889.37	--

Week A is exactly the issue-#2 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 71.1% is the audited daily-candle figure -- different estimators, both reported as measured.

Practical implications

- Do not read this week's leaderboard as a skill ranking: no title was awarded, the top CIs overlap, and the two biggest jumps (+17.5pp, +14.7pp) came from the two models that were last and second-to-last a week earlier.
- Treat per-ticker readability as weekly weather. Three issues, three different best tickers (SOL -> BNB -> BTC) and the two-week 'ETH is hardest' pattern broke.
- Self-agreement stopped being a penalty but is not yet a booster: pooled lifts -0.6 to +4.6pp with overlapping CIs, and the reversal result is two events. Carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Aug 17-23, 2026 UTC). Descriptive for this window only; no claim about next week or any model's underlying skill. This window's regime is the opposite of issue #2's: 5 trend days of 7 against 1 of 7.
- 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. Four events across three issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are still small (n=5, 11, 41, 42); the n=5 cell is 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.
- qwen-3.8-max ran a 96.20% ok-rate this window; its mature pool is 1,281 against 1,330 for the other six models. Models in this window: claude-fable-5, claude-opus-5, deepseek-v4-pro, gemini-3.1-pro, gpt-5.6-sol, grok-4.5, qwen-3.8-max. Series density and lineage notes (wave 3): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 17-23) daily coverage is PARTIAL by design: the 1w series has the Mon Aug 17 anchor plus daily slots on Aug 22-23 only, and the 1M series has a daily slot on Aug 23 only; both sit outside this report's FH gate (1h/4h/1d) and appear only in the exclusions counter (1w 209, 1M 68). (2) After the report window — from Aug 24, 2026 — the grok line runs Grok 4.6; every grok forecast in this window and in the week-2 comparison is grok-4.5. (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 this issue, after an audit found the raw table mixes two exchanges; issue #2 row is as published.
- Research-to-date counter pinned from this issue: directional forecasts resolved inside the FH gate (1h/4h/1d) since Jul 11, counted at the issue cutoff (Aug 24 16:00 UTC). Issue #2 printed 23,458 under an earlier, unpinned definition; treat cross-issue counter deltas across the pin as definitional, not additive.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	9,590	Mature (scored pool)	9,260
OK in gate	9,260	-- of them directional	5,130
Out of gate (1w / 1M)	209 / 68	-- of them sideways	4,130
Invalid	53	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-08-17.json	Report cutoff	Mon Aug 24, 2026, 16:00 UTC
Generated at (pipeline)	2026-08-26 05:22 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,130 scored observations. MARKETMANIA RESEARCH TO DATE (as of Aug 24, 2026 cutoff): 33,809 directional forecasts resolved since Jul 11 · 11 models tracked (7 current + 4 archived legacy) · 5 assets · 5 horizons · hourly · 13 published reports.

Issue #3. Weekly Model Watch is a living series. Issue #2's open questions -- would a third reversal repeat the blind spot, does SOL's readability bounce back, and does the trend-day agreement premium reappear? -- closed 'no' (both ETH turns found callers), 'yes' (38.9% -> 54.3%) and 'yes' (64.8% trend vs 47.5% flat on 4h TF-agreement, now on a readable split). After the report window — from Aug 24, 2026 — the grok line runs Grok 4.6; every grok forecast in this window and in the week-2 comparison is grok-4.5. Engine 1.1 has powered the sandbox since Aug 18, i.e. from day 2 of this window; no cross-engine PnL comparisons are claimed.

What we're testing next

- Next issue: does the leaderboard reshuffle again when the tape calms, or is claude-opus-5's +17.5pp the start of something? Does the reversal detector find callers on a non-ETH turn?
- Does the every-ticker-above-51% floor survive a flat week?
- Monthly test (Sep 2): cross-TF/FH lifts with regime control across mixed weeks (trend vs. flat), block bootstrap.

Related research

Report	Direct PDF link
Consensus Watch #3	https://marketmania.ai/research/reports/consensus-watch-2026-08-17.pdf
Weekly Calibration #3	https://marketmania.ai/research/reports/weekly-calibration-2026-08-17.pdf
Weekly Model Watch #2	https://marketmania.ai/research/reports/model-watch-2026-08-10.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_2026w34,
  title = {Weekly Model Watch #3: weekly leaderboard, ticker/pair reads and cross-confirmation, Aug 17-23 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {August}, day = {26},
  url   = {https://marketmania.ai/research/reports/model-watch-2026-08-17.pdf},
  note  = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Aug 17-23, 2026 UTC; source weekly_metrics_2026-08-17.json}
}
```