

Weekly Model Watch #6

September 15, 2026 · MarketMania Research · Weekly series · window Sep 7-13, 2026 UTC

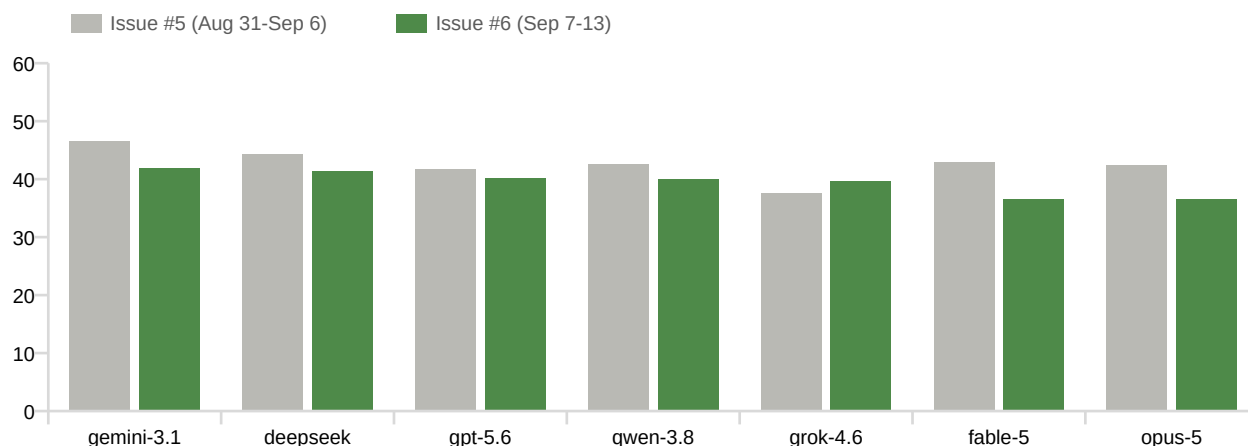
RESEARCH SNAPSHOT			
4,468 directional calls scored	7 models (stable lineup)	of 9,301 mature forecasts	window Sep 7-13, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 39.6%	cutoff Mon Sep 14, 2026, 16:00 UTC
market: BTC net -4.36% (prior +3.42%)	ann vol 19.7% (was 43.4%)	TOP5 volume \$16.05B , -1% w/w	pairwise corr 0.74 (was 0.78)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 42.8% -> 39.6% (-3.3pp) with 1 of 7 models improving, and no weekly title was awarded, for the sixth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** gemini-3.1-pro leads descriptively at 41.8% [38.1%, 45.5%]; the gap to #2 (deepseek-v4-pro, 41.4%) is 0.4pp. 1 of 7 models improved week over week; issue #5's leader gemini-3.1-pro is now 1 of 7.
- **No ticker cleared 51%.** BNB is the week's most readable coin (46.9%, n=963) and ETH the hardest at 29.5%; 0 of the 5 assets cleared 51%. Pair of the week: gemini-3.1-pro x BNB 50.0% (n=148); worst: grok-4.6 x ETH 25.6% (n=82).
- **1 of 2 qualifying turns found a caller.** ETH long +4.19%, 11 caller(s), first gemini-3.1-pro; ETH short -2.11%, 0 caller(s), first NOBODY.
- **Self-agreement split across the four cells.** Cross-TF/FH lifts ran -6.0pp to +0.3pp (issue #5: -17.4pp to -2.5pp). On trend days, 4h TF-agreement hit 52.8% (n=195) against 21.2% (n=325) 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 42.8% -> 39.6%. The grok column compares two pure grok-4.6 weeks: the 4.5 -> 4.6 flip sits two windows back. "Was" values are the numbers published in issue #5; 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. The grok row is a pure grok-4.6 week in this issue and was one in issue #5 as well -- the 4.5 -> 4.6 flip sits two windows back -- 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 two windows back (issue #4): both weeks of this issue are pure grok-4.6, with no 4.5-era rows in either window, so the grok week-over-week row is the first pure-4.6 against pure-4.6 comparison of the series.

Leaderboard of the week

Model	N	Cov.	Hit rate	95% CI	W/w pp
gemini-3.1-pro	673	50.6%	41.8%	38.1%-45.5%	-4.8
deepseek-v4-pro	727	54.7%	41.4%	37.9%-45.0%	-2.8
gpt-5.6-sol	709	53.4%	40.1%	36.5%-43.7%	-1.6
qwen-3.8-max	707	53.4%	40.0%	36.5%-43.7%	-2.4
grok-4.6	518	39.0%	39.6%	35.5%-43.9%	+2.0
claude-fable-5	624	46.9%	36.5%	32.9%-40.4%	-6.3
claude-opus-5	510	38.4%	36.5%	32.4%-40.7%	-5.9

No weekly title is awarded this issue -- for the sixth 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 0.4pp and their CIs overlap, so neither condition is met. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits two windows back (issue #4): both weeks of this issue are pure grok-4.6, with no 4.5-era rows in either window, so the grok week-over-week row is the first pure-4.6 against pure-4.6 comparison of the series.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #5
BNB	963	46.9%	43.8%-50.1%	47.3%
XRP	937	42.2%	39.0%-45.3%	44.2%
BTC	891	41.4%	38.2%-44.7%	40.6%
SOL	875	36.0%	32.9%-39.2%	38.7%
ETH	802	29.5%	26.5%-32.8%	43.1%

BNB was the field's most readable ticker (46.9%) and ETH the hardest (29.5%); 0 of 5 assets cleared 51% this week (issue #5: 0 of 5). Issue #5's best ticker was BNB and its hardest SOL -- this issue's ordering is BNB > XRP > BTC > SOL > ETH.

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
gemini-3.1-pro	BNB	148	50.0%
deepseek-v4-pro	BNB	162	49.4%
grok-4.6	BNB	112	49.1%
gpt-5.6-sol	BNB	157	48.4%
claude-opus-5	BNB	118	45.8%
deepseek-v4-pro	BTC	153	45.1%

Model (5 worst)	Symbol	N	Hit rate
gemini-3.1-pro	ETH	122	30.3%
claude-opus-5	ETH	95	29.5%
gpt-5.6-sol	ETH	121	28.1%
deepseek-v4-pro	ETH	132	28.0%
grok-4.6	ETH	82	25.6%

gemini-3.1-pro x BNB (50.0%, n=148) is the week's best pair and grok-4.6 x ETH (25.6%, n=82) the weakest. Pair history accumulates across issues -- read this as the sixth 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	Fri Sep 11, 08:00	Long	+4.19%	11	gemini-3.1-pro
ETH	Fri Sep 11, 16:00	Short	-2.11%	0	NOBODY

The ETH long turn drew 11 caller(s), first gemini-3.1-pro on a 1d/1d call at Sep 12, 00:01 with confidence 70; the ETH short turn drew NOBODY. Scope in v1 remains BTC/ETH only; the event count across six 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	520	33.1%	10	80.0%	35.4% (704)	-2.3
cross-TF	1d	TF 1d vs 4h	84	11.9%	5 *	40.0%	16.7% (132)	-4.8
cross-FH	4h	by FH 1h	485	35.7%	6 *	83.3%	35.4% (704)	+0.3
cross-FH	1d	by FH 4h	75	10.7%	8 *	12.5%	16.7% (132)	-6.0

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 173 / 43 / 211 / 49. Pooled across all four cell families the lifts run -6.0pp to +0.3pp this week (issue #5: -17.4pp to -2.5pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has 2 trend days of 7: 09-07 -1.52% (trend), 09-08 -0.72%, 09-09 -0.25%, 09-10 -2.14% (trend), 09-11 +0.85%, 09-12 +0.00%, 09-13 -0.57%. On trend days, 4h TF-agreement hit 52.8% (n=195) against 21.2% (n=325) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	20.7%	26.2%	53.1%	0.79
claude-opus-5	1,330	17.9%	20.4%	61.7%	0.87
deepseek-v4-pro	1,330	19.9%	34.7%	45.3%	0.57
gemini-3.1-pro	1,330	19.3%	31.3%	49.4%	0.62
gpt-5.6-sol	1,327	16.4%	37.0%	46.6%	0.44
grok-4.6	1,330	11.4%	27.5%	61.1%	0.42
qwen-3.8-max	1,324	20.5%	32.9%	46.6%	0.63

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 0.62 this week. 'Sideways' was 52.0% of mature forecasts (4,833 of 9,301). Consensus skew: 533 symbol/FH/TF cells had 5 or more models on the same side and hit 38.1% (n=3,347) against the 39.6% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gpt-5.6-sol	BTC	Short	82	1h / 1h	Sep 8, 07:01	EXPIRY	-\$0.23

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

The hit-rate leader this week is gemini-3.1-pro (41.8%); the best-calibrated line is qwen-3.8-max (Brier 0.2794, gap +18.8pp). Per-model ok-rate this window ran 99.59% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #6.

Market check: a down week on a quieter tape

OBSERVATION — mean |1d move| 2.10% -> 1.35% (-36% rel), BTC realized vol 35.2% -> 31.4% (ann., hourly); field directional accuracy 42.8% -> 39.6% (-3.3pp), 1 of 7 models improved, sim win-rate up for 1 of 7, field sim PnL -\$669.03 -> -\$1018.93 (adjacent calendar weeks Aug 31-Sep 6 vs Sep 7-13; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 51.6% -> 38.7% — the same direction as the trade-based hit rule.

Measure	Aug 31-Sep 6 (week A)	Sep 7-13 (week B)	Change
Mean 1d move (5 assets)	2.10%	1.35%	-36% rel
BTC realized vol (ann., hourly)	35.2%	31.4%	-3.8pp
Field directional accuracy	42.8%	39.6%	-3.3pp
Models improving hit-rate	--	1 of 7	--
Raw price-sign accuracy	51.6%	38.7%	-12.8pp
Field sim win-rate	35.6%	32.1%	-3.5pp
Field sim net PnL	-\$669.03	-\$1018.93	--

Week A is exactly the issue-#5 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, two windows back, inside issue #4's window). Week A (2026-08-31..2026-09-06) is pure grok-4.6 (it reproduces the published issue #5); 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. Six issues, and the best ticker changed hands in every one until this week (SOL -> BNB -> BTC -> SOL -> BNB -> BNB): BNB led issue #5 and leads again at 46.9%.
- Self-agreement lifts run -6.0pp to +0.3pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Sep 7-13, 2026 UTC). Descriptive for this window only; no claim about next week or any model's underlying skill. This window ran 2 trend days of 7 (issue #5: 3 of 7); every comparison with issue #5 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. 12 events across six issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=10, 5, 6, 8); 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 6): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 7-13) daily coverage is FULL, as it was in issues #4 and #5: 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, two windows back (inside the issue-#4 window): this window carries 1,470 grok rows and none from the 4.5 era, and neither does week A (the issue-#5 window), so every grok number in this issue and in issue #5 is a pure grok-4.6 line -- the grok week-over-week row is the first pure-4.6 against pure-4.6 comparison of the series. (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 #5 after the audit that found the raw table mixes two exchanges; issue #5 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 14 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4 and #5, which printed 42,905 at the Sep 7 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 47,373 is an additive step under one definition; counter deltas against issue #4 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	4,468
Out of gate (1w / 1M)	490 / 490	-- of them sideways	4,833
Invalid	9	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-07.json	Report cutoff	Mon Sep 14, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-15 15:24 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,468 scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 14, 2026 cutoff): 47,373 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 29 published reports.

Issue #6. Weekly Model Watch is a living series. Issue #5'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 at all, on a non-ETH turn? -- read this issue as: gemini-3.1-pro at the top (41.8%), the line that also led issue #5 (gemini-3.1-pro), so the rank held two issues running; no weekly title was awarded, for the sixth issue running, and 1 of 2 qualifying turns found a caller, on turns in ETH only. The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #5; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits two windows back. 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 gemini-3.1-pro hold the top rank for a third issue, and does the title rule ever fire?
- Does a qualifying reversal ever fire outside ETH, and does a second turn in the same pair ever draw a caller?
- 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 #6	https://marketmania.ai/research/reports/consensus-watch-2026-09-07.pdf
Weekly Calibration #6	https://marketmania.ai/research/reports/weekly-calibration-2026-09-07.pdf
Weekly Model Watch #5	https://marketmania.ai/research/reports/model-watch-2026-08-31.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_2026w37,
  title = {Weekly Model Watch #6: weekly leaderboard, ticker/pair reads and cross-confirmation, Sep 7-13 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {15},
  url = {https://marketmania.ai/research/reports/model-watch-2026-09-07.pdf},
  note = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Sep 7-13, 2026 UTC; source weekly_metrics_2026-09-07.json}
}
```