Weekly Model Watch #14

November 11, 2026 · MarketMania Research · Weekly series · window Nov 2-8, 2026 UTC

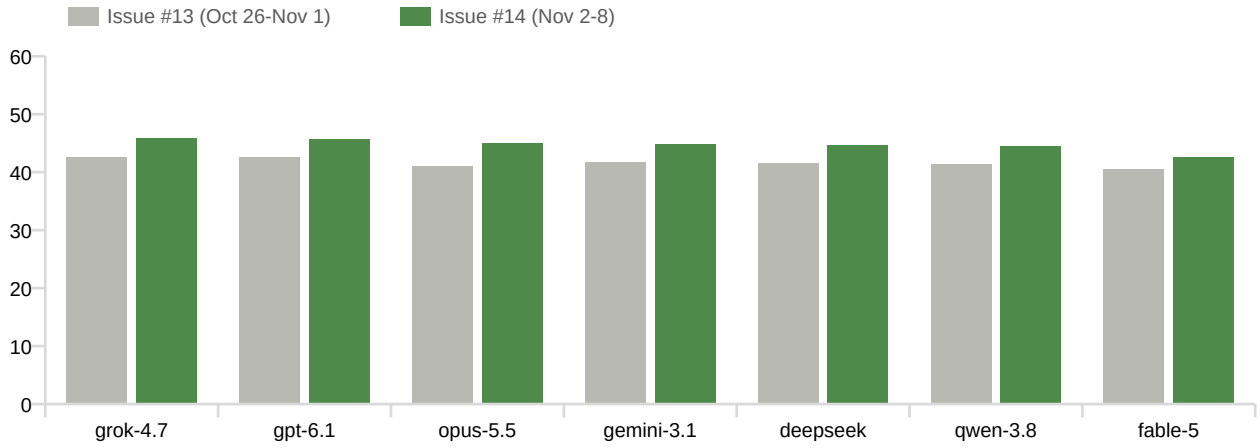
RESEARCH SNAPSHOT			
4,865 directional calls scored	**7** models (stable lineup)	of **9,250** mature forecasts	window **Nov 2-8, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **44.8%**	cutoff **Mon Nov 9, 2026, 16:00 UTC**
market: BTC net **-4.47%** (prior -2.25%)	ann vol **55.3%** (was 34.7%)	TOP5 volume **$14.26B**, -4% w/w	pairwise corr **0.94** (was 0.93)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 41.6% -> 44.8% (+3.1pp) with 7 of 7 models improving, and no weekly title was awarded, for the fourteenth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** grok-4.7 leads descriptively at 45.9% [42.4%, 49.4%]; the gap to #2 (gpt-6.1-sol, 45.6%) is 0.3pp. 7 of 7 models improved week over week; issue #13's leader grok-4.7 is now 1 of 7.
- **1 of 5 tickers were readable.** XRP is the week's most readable coin (51.7%, n=992) and SOL the hardest at 38.4%; 1 of the 5 assets cleared 51%. Pair of the week: gpt-6.1-sol x XRP 55.6% (n=169); worst: claude-fable-5 x BNB 31.9% (n=116).
- **2 of 5 qualifying turns found a caller.** BTC long +2.54%, 0 caller(s), first NOBODY; BTC short -5.49%, 14 caller(s), first gemini-3.1-pro; ETH long +3.43%, 0 caller(s), first NOBODY; ETH short -7.06%, 11 caller(s), first gpt-6.1-sol; ETH short -5.50%, 0 caller(s), first NOBODY.
- **Self-agreement read as a penalty in every cell.** Cross-TF/FH lifts ran -8.2pp to -2.1pp (issue #13: -4.0pp to +3.9pp). On trend days, 4h TF-agreement hit 49.7% (n=366) against 37.1% (n=221) on flat days.



Directional hit-rate (%) by model, week over week, ranked by this week's rate. Short names: opus-5.5 = claude-opus-5.5, fable-5 = claude-fable-5, gemini-3.1 = gemini-3.1-pro, gpt-6.1 = gpt-6.1-sol, qwen-3.8 = qwen-3.8-max, deepseek = deepseek-v4-pro. Field base 41.6% -> 44.8%. "Was" values are the numbers published in issue #13; 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 claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are single-build weeks on the Oct 17 builds in both this issue and issue #13, so those three week-over-week rows compare like with like.

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 >=5pp gap AND non-overlapping 95% Wilson CIs with N>=10 per side -- a descriptive lead is not a title. Reversals use the 4h grid with a >=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
grok-4.7	771	58.1%	45.9%	42.4%-49.4%	+3.4
gpt-6.1-sol	807	60.7%	45.6%	42.2%-49.0%	+3.1
claude-opus-5.5	516	38.8%	45.0%	40.7%-49.3%	+4.0
gemini-3.1-pro	752	56.6%	44.8%	41.3%-48.4%	+3.1
deepseek-v4-pro	754	57.1%	44.6%	41.0%-48.1%	+3.1
qwen-3.8-max	663	51.6%	44.5%	40.8%-48.3%	+3.1
claude-fable-5	602	45.3%	42.5%	38.6%-46.5%	+2.1

No weekly title is awarded this issue -- for the fourteenth time. Title rule: **>=5pp gap AND non-overlapping 95% CI, N>=10** per side. grok-4.7's lead over gpt-6.1-sol is 0.3pp and their CIs overlap, so neither condition is met.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #13
XRP	992	51.7%	48.6%-54.8%	42.7%
BTC	1,049	47.8%	44.8%-50.8%	42.3%
ETH	966	45.6%	42.5%-48.8%	41.8%
BNB	858	39.5%	36.3%-42.8%	36.8%
SOL	1,000	38.4%	35.4%-41.4%	44.0%

XRP was the field's most readable ticker (51.7%) and SOL the hardest (38.4%); 1 of 5 assets cleared 51% this week (issue #13: 0 of 5). Issue #13's best ticker was SOL and its hardest BNB -- this issue's ordering is XRP > BTC > ETH > BNB > 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
gpt-6.1-sol	XRP	169	55.6%
deepseek-v4-pro	XRP	161	52.8%
claude-opus-5.5	XRP	104	51.9%
claude-fable-5	XRP	116	51.7%
grok-4.7	XRP	156	51.3%
grok-4.7	BTC	171	50.9%

Model (5 worst)	Symbol	N	Hit rate
deepseek-v4-pro	SOL	144	38.2%
claude-opus-5.5	SOL	103	37.9%
grok-4.7	SOL	151	37.1%
qwen-3.8-max	SOL	134	34.3%
claude-fable-5	BNB	116	31.9%

gpt-6.1-sol x XRP (55.6%, n=169) is the week's best pair and claude-fable-5 x BNB (31.9%, n=116) the weakest. Pair history accumulates across issues -- read this as the fourteenth 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	Wed Nov 4, 04:00	Long	+2.54%	0	NOBODY
BTC	Thu Nov 5, 08:00	Short	-5.49%	14	gemini-3.1-pro
ETH	Wed Nov 4, 04:00	Long	+3.43%	0	NOBODY
ETH	Thu Nov 5, 08:00	Short	-7.06%	11	gpt-6.1-sol
ETH	Sun Nov 8, 16:00	Short	-5.50%	0	NOBODY

The BTC long turn drew NOBODY; the BTC short turn drew 14 caller(s), first gemini-3.1-pro on a 1d/4h call at Nov 6, 00:01 with confidence 75; the ETH long turn drew NOBODY; the ETH short turn drew 11 caller(s), first gpt-6.1-sol on a 1d/4h call at Nov 6, 00:01 with confidence 73; the ETH short turn drew NOBODY. Scope in v1 remains BTC/ETH only; the event count across fourteen 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	587	45.0%	14	57.1%	47.0% (882)	-2.1
cross-TF	1d	TF 1d vs 4h	121	48.8%	8 *	62.5%	51.8% (166)	-3.0
cross-FH	4h	by FH 1h	505	44.2%	35	60.0%	47.0% (882)	-2.9
cross-FH	1d	by FH 4h	101	43.6%	8 *	62.5%	51.8% (166)	-8.2

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 276 / 37 / 338 / 57. Pooled across all four cell families the lifts run -8.2pp to -2.1pp this week (issue #13: -4.0pp to +3.9pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has trend on 4 of 7 days: 11-02 -2.04% (trend), 11-03 +0.92%, 11-04 +3.90% (trend), 11-05 -5.33% (trend), 11-06 -0.47%, 11-07 -0.07%, 11-08 -1.60% (trend). On trend days, 4h TF-agreement hit 49.7% (n=366) against 37.1% (n=221) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	25.0%	20.2%	54.7%	1.24
claude-opus-5.5	1,330	18.8%	20.0%	61.2%	0.94
deepseek-v4-pro	1,320	22.4%	34.7%	42.9%	0.65
gemini-3.1-pro	1,329	24.8%	31.8%	43.4%	0.78
gpt-6.1-sol	1,330	15.9%	44.8%	39.3%	0.35
grok-4.7	1,327	22.7%	35.4%	41.9%	0.64
qwen-3.8-max	1,284	23.4%	28.3%	48.4%	0.83

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 0.77 this week. 'Sideways' was 47.4% of mature forecasts (4,385 of 9,250). Consensus skew: 532 symbol/FH/TF cells had 5 or more models on the same side and hit 44.2% (n=3,390) against the 44.8% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	BTC	Long	80	1d / 4h	Nov 5, 00:01	SL	-$2.60

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 grok-4.7 (45.9%); the lowest-Brier line is claude-opus-5.5 (Brier 0.2609, gap +11.3pp). Per-model ok-rate this window ran 96.87% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #14.

Market check: bigger daily moves on lower volume

OBSERVATION — mean |1d move| 1.63% -> 2.16% (+33% rel), BTC realized vol 37.0% -> 41.0% (ann., hourly); field directional accuracy 41.6% -> 44.8% (+3.1pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$869.65 -> -$464.89 (adjacent calendar weeks Oct 26-Nov 1 vs Nov 2-8; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 42.8% -> 48.9% — the same direction as the trade-based hit rule.

Measure	Oct 26-Nov 1 (week A)	Nov 2-8 (week B)	Change
Mean |1d move| (5 assets)	1.63%	2.16%	+33% rel
BTC realized vol (ann., hourly)	37.0%	41.0%	+4.0pp
Field directional accuracy	41.6%	44.8%	+3.1pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	42.8%	48.9%	+6.0pp
Field sim win-rate	33.6%	37.4%	+3.8pp
Field sim net PnL	-$869.65	-$464.89	--

Week A is exactly the issue-#13 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 note: the model cutover (2026-10-17T09:01:00+00:00, day 6 of 7 of its Mon-Sun week) sits three windows back, inside the published issue-#11 window, where it spliced three lines -- gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5, grok-4.6 -> grok-4.7. Week A (2026-10-26..2026-11-01, the published issue-#13 window) and week B (2026-11-02..2026-11-08) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #13. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits ten windows back (inside issue #4's window).

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. Fourteen issues, and the best ticker changed hands in every one except issues #6, #10 and #12 (SOL -> BNB -> BTC -> SOL -> BNB -> BNB -> ETH -> XRP -> BTC -> BTC -> XRP -> XRP -> SOL -> XRP): XRP leads this week at 51.7%, after SOL in issue #13.
- Self-agreement lifts run -8.2pp to -2.1pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Nov 2-8, 2026 UTC). Descriptive for this window only; no claim about next week or any model's underlying skill. This window had trend on 4 of 7 days (issue #13: 1 of 7); every comparison with issue #13 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. 37 events across fourteen issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=14, 8, 35, 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-5.5, deepseek-v4-pro, gemini-3.1-pro, gpt-6.1-sol, grok-4.7, qwen-3.8-max. DeepSeek announced routing of V4 Pro to V4.1 Flash from Sep 14, 04:00 UTC; an API check on Nov 10 still returns deepseek-v4-pro as the served model for our requests, with the same system

fingerprint as on Nov 3, so this issue keeps the label; it will change the day the served id changes. In this window the line's ok-rate is 99.9% (invalid 1 of its 1,461 rows; 54 across all lines this window). Provider incident: on Nov 5 the provider behind qwen-3.8-max rejected 46 of its 210 first-attempt forecasts with an input-moderation error; those rows are counted in Invalid. From Nov 6 the engine retries such a call without the news block; 2 forecasts in this window (qwen-3.8-max: 2 on Nov 7) were produced that way and are flagged `news_stripped` in the open data. Hit rates in this report use the rows that resolved. Series density and lineage notes (wave 14): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Nov 2-8) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11, #12 and #13: 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 487). (2) This window is the third full week on the builds that took over at Oct 17, 2026 09:01 UTC: claude-opus-5.5, gpt-6.1-sol and grok-4.7 carry no rows from the old builds in either week, and the 7 lines carry 1,461 / 1,470 rows in this window and 1,470 in week A, so every week-over-week row compares two full weeks on the same builds. (3) No roll-out slot of the flip falls inside either week: the two hourly slots before it (Oct 17, 07:01 and 08:01 UTC) lie three windows back, so this window's invalid counter holds none of their rows. (4) Past lineage flips: the Oct 17, 2026 09:01 UTC flip sits three windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits ten windows back (inside the issue-#4 window), and 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 #13 after the audit that found the raw table mixes two exchanges; issue #13 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 (Nov 9 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11, #12 and #13, which printed 82,895 at the Nov 2 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 87,760 is an additive step under one definition; counter deltas against issue #12 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,281	Mature (scored pool)	9,250
OK in gate	9,250	-- of them directional	4,865
Out of gate (1w / 1M)	490 / 487	-- of them sideways	4,385
Invalid	54	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-11-02.json	Report cutoff	Mon Nov 9, 2026, 16:00 UTC
Generated at (pipeline)	2026-11-10 14:05 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,865 scored observations. MARKETMANIA RESEARCH TO DATE (as of Nov 9, 2026 cutoff): 87,760 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 73 published reports.

Issue #14. Weekly Model Watch is a living series. Issue #13's open questions -- "does grok-4.7 hold the top rank for a second issue, and does the title rule ever fire?" and "Does a qualifying reversal fire outside ETH for a fourth issue running, and does a second turn in the same pair draw a caller?" -- read this issue as: grok-4.7 at the top (45.9%), the same line that led issue #13 (grok-4.7), so the rank held for a second issue; no weekly title was awarded, for the fourteenth issue running, and 2 of 5 qualifying turns found a caller, on turns in BTC / ETH; the two BTC turns fired outside ETH for a fourth issue running, and the second turn in each pair (Thu Nov 5, 08:00) drew callers -- 14 on BTC and 11 on ETH -- while the third ETH turn (Sun Nov 8, 16:00) drew NOBODY. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the third full week on the Oct 17 builds; week-over-week rows compare two full weeks on the same builds. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits ten 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 grok-4.7 hold the top rank for a third issue, and does the title rule ever fire?
- Does a qualifying reversal fire outside ETH for a fifth issue running, and does a second turn in the same pair draw a caller again?
- 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 #14	https://marketmania.ai/research/reports/consensus-watch-2026-11-02.pdf
Weekly Calibration #14	https://marketmania.ai/research/reports/weekly-calibration-2026-11-02.pdf
Weekly Model Watch #13	https://marketmania.ai/research/reports/model-watch-2026-10-26.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_2026w45,
  title  = {Weekly Model Watch #14: weekly leaderboard, ticker/pair reads and cross-confirmation, Nov 2-8 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {11},
  url    = {https://marketmania.ai/research/reports/model-watch-2026-11-02.pdf},
  note   = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Nov 2-8, 2026 UTC; source weekly_metrics_2026-11-02.json}
}
```