Weekly Model Watch #15

November 18, 2026 · MarketMania Research · Weekly series · window Nov 9-15, 2026 UTC

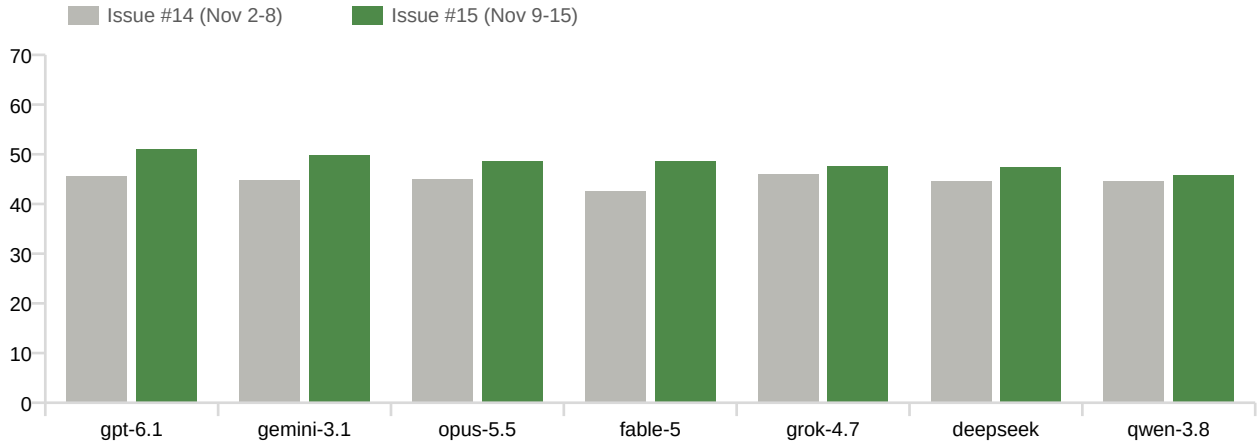
RESEARCH SNAPSHOT			
5,669 directional calls scored	**7** models (stable lineup)	of **9,298** mature forecasts	window **Nov 9-15, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **48.5%**	cutoff **Mon Nov 16, 2026, 16:00 UTC**
market: BTC net **-12.69%** (prior -4.47%)	ann vol **57.5%** (was 55.3%)	TOP5 volume **$18.77B**, +32% w/w	pairwise corr **0.95** (was 0.94)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 44.8% -> 48.5% (+3.7pp) with 7 of 7 models improving, and no weekly title was awarded, for the fifteenth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** gpt-6.1-sol leads descriptively at 51.1% [48.0%, 54.1%]; the gap to #2 (gemini-3.1-pro, 49.8%) is 1.3pp. 7 of 7 models improved week over week; issue #14's leader grok-4.7 is now 5 of 7.
- **No ticker cleared 51%.** SOL is the week's most readable coin (50.9%, n=1,185) and BTC the hardest at 46.1%; 0 of the 5 assets cleared 51%. Pair of the week: gpt-6.1-sol x SOL 55.6% (n=214); worst: qwen-3.8-max x BTC 41.5% (n=147).
- **3 of 4 qualifying turns found a caller.** BTC long +2.27%, 0 caller(s), first NOBODY; BTC short -2.09%, 14 caller(s), first gpt-6.1-sol; ETH long +2.65%, 1 caller(s), first gemini-3.1-pro; ETH short -4.11%, 29 caller(s), first gemini-3.1-pro.
- **Self-agreement read as a penalty in every cell.** Cross-TF/FH lifts ran -2.0pp to -1.1pp (issue #14: -8.2pp to -2.1pp). On trend days, 4h TF-agreement hit 49.8% (n=540) against 17.4% (n=115) 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 44.8% -> 48.5%. "Was" values are the numbers published in issue #14; 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 #14, 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
gpt-6.1-sol	1,034	77.7%	51.1%	48.0%-54.1%	+5.5
gemini-3.1-pro	790	59.5%	49.8%	46.3%-53.2%	+4.9
claude-opus-5.5	613	46.1%	48.6%	44.7%-52.6%	+3.6
claude-fable-5	705	53.0%	48.5%	44.8%-52.2%	+6.0
grok-4.7	963	72.9%	47.6%	44.4%-50.7%	+1.7
deepseek-v4-pro	839	63.1%	47.4%	44.1%-50.8%	+2.9
qwen-3.8-max	725	54.5%	45.7%	42.1%-49.3%	+1.2

No weekly title is awarded this issue -- for the fifteenth time. Title rule: **>=5pp gap AND non-overlapping 95% CI, N>=10** per side. gpt-6.1-sol's lead over gemini-3.1-pro is 1.3pp and their CIs overlap, so neither condition is met.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #14
SOL	1,185	50.9%	48.0%-53.7%	38.4%
BNB	1,075	49.5%	46.5%-52.5%	39.5%
XRP	1,117	48.3%	45.4%-51.3%	51.7%
ETH	1,134	47.5%	44.6%-50.4%	45.6%
BTC	1,158	46.1%	43.3%-49.0%	47.8%

SOL was the field's most readable ticker (50.9%) and BTC the hardest (46.1%); 0 of 5 assets cleared 51% this week (issue #14: 1 of 5). Issue #14's best ticker was XRP and its hardest SOL -- this issue's ordering is SOL > BNB > XRP > ETH > BTC.

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	SOL	214	55.6%
gpt-6.1-sol	BNB	188	51.6%
gemini-3.1-pro	SOL	165	51.5%
claude-opus-5.5	BNB	111	51.3%
claude-fable-5	SOL	150	51.3%
gemini-3.1-pro	BNB	150	51.3%

Model (5 worst)	Symbol	N	Hit rate
deepseek-v4-pro	ETH	175	44.6%
claude-opus-5.5	BTC	130	43.9%
qwen-3.8-max	XRP	147	43.5%
claude-fable-5	BTC	152	43.4%
qwen-3.8-max	BTC	147	41.5%

gpt-6.1-sol x SOL (55.6%, n=214) is the week's best pair and qwen-3.8-max x BTC (41.5%, n=147) the weakest. Pair history accumulates across issues -- read this as the fifteenth 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	Mon Nov 9, 08:00	Long	+2.27%	0	NOBODY
BTC	Tue Nov 10, 08:00	Short	-2.09%	14	gpt-6.1-sol
ETH	Mon Nov 9, 08:00	Long	+2.65%	1	gemini-3.1-pro
ETH	Tue Nov 10, 08:00	Short	-4.11%	29	gemini-3.1-pro

The BTC long turn drew NOBODY; the BTC short turn drew 14 caller(s), first gpt-6.1-sol on a 1d/4h call at Nov 10, 00:01 with confidence 61; the ETH long turn drew 1 caller(s), first gemini-3.1-pro on a 4h/4h call at Nov 9, 08:01 with confidence 60; the ETH short turn drew 29 caller(s), first gemini-3.1-pro on a 1d/1d call at Nov 10, 00:01 with confidence 70. Scope in v1 remains BTC/ETH only; the event count across fifteen 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	655	44.1%	51	66.7%	46.2% (1,057)	-2.0
cross-TF	1d	TF 1d vs 4h	177	66.7%	3 *	33.3%	67.8% (208)	-1.1
cross-FH	4h	by FH 1h	595	44.9%	67	62.7%	46.2% (1,057)	-1.3
cross-FH	1d	by FH 4h	156	66.7%	1 *	100.0%	67.8% (208)	-1.1

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 350 / 28 / 394 / 50. Pooled across all four cell families the lifts run -2.0pp to -1.1pp this week (issue #14: -8.2pp to -2.1pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has trend on 5 of 7 days: 11-09 +0.93%, 11-10 -2.06% (trend), 11-11 +0.46%, 11-12 -3.58% (trend), 11-13 -4.24% (trend), 11-14 -6.26% (trend), 11-15 +1.72% (trend). On trend days, 4h TF-agreement hit 49.8% (n=540) against 17.4% (n=115) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	12.8%	40.2%	47.0%	0.32
claude-opus-5.5	1,330	9.8%	36.3%	53.9%	0.27
deepseek-v4-pro	1,330	8.4%	54.7%	36.9%	0.15
gemini-3.1-pro	1,327	13.2%	46.4%	40.5%	0.28
gpt-6.1-sol	1,330	6.1%	71.7%	22.3%	0.08
grok-4.7	1,321	13.8%	59.1%	27.1%	0.23
qwen-3.8-max	1,330	10.0%	44.5%	45.5%	0.22

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 0.22 this week. 'Sideways' was 39.0% of mature forecasts (3,629 of 9,298). Consensus skew: 652 symbol/FH/TF cells had 5 or more models on the same side and hit 48.0% (n=4,070) against the 48.5% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	SOL	Short	80	4h / 4h	Nov 9, 08:01	SL	-$2.10

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 gpt-6.1-sol (51.1%); the lowest-Brier line is claude-opus-5.5 (Brier 0.2555, gap +7.2pp). Per-model ok-rate this window ran 99.12% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #15.

Market check: bigger daily moves on higher volume

OBSERVATION — mean |1d move| 2.16% -> 3.40% (+57% rel), BTC realized vol 41.0% -> 55.8% (ann., hourly); field directional accuracy 44.8% -> 48.5% (+3.7pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$464.89 -> +$160.79 (adjacent calendar weeks Nov 2-8 vs Nov 9-15; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 48.9% -> 54.4% — the same direction as the trade-based hit rule.

Measure	Nov 2-8 (week A)	Nov 9-15 (week B)	Change
Mean |1d move| (5 assets)	2.16%	3.40%	+57% rel
BTC realized vol (ann., hourly)	41.0%	55.8%	+14.8pp
Field directional accuracy	44.8%	48.5%	+3.7pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	48.9%	54.4%	+5.6pp
Field sim win-rate	37.4%	43.9%	+6.4pp
Field sim net PnL	-$464.89	+$160.79	--

Week A is exactly the issue-#14 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 four 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 (2026-11-02..2026-11-08, the published issue-#14 window) and week B (2026-11-09..2026-11-15) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #14. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits eleven 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. Fifteen 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 -> SOL): SOL leads this week at 50.9%, after XRP in issue #14.
- Self-agreement lifts run -2.0pp to -1.1pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Nov 9-15, 2026 UTC). Descriptive for this window only; no claim about next week or any model's underlying skill. This window had trend on 5 of 7 days (issue #14: 4 of 7); every comparison with issue #14 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. 41 events across fifteen issues cannot characterise detector performance either way.

- Cross-TF and cross-FH disagreement cells are small (n=51, 3, 67, 1); 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 18 still returns deepseek-v4-pro as the served model for our requests, with a system fingerprint that differs from Nov 10, so this issue keeps the label; it will change the day the served id changes. In this window the line's ok-rate is 100.0% (invalid 0 of its 1,470 rows; 16 across all lines this window). Provider note: grok-4.7 returned no valid answer for 13 of its 1,470 forecasts in this window (8 on Nov 9, 3 on Nov 10, 1 on Nov 11, 1 on Nov 15); those rows are counted in Invalid. The engine retries a call the provider rejects on input moderation without the news block (since Nov 6); 4 forecasts in this window (qwen-3.8-max: 1 on Nov 10, 2 on Nov 11, 1 on Nov 13) 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 15): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Nov 9-15) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11, #12, #13 and #14: 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 489, 1M 487). (2) This window is the fourth 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 1,470 rows in this window and 1,461 / 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 four 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 four windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eleven 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 #14 after the audit that found the raw table mixes two exchanges; issue #14 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 16 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11, #12, #13 and #14, which printed 87,760 at the Nov 9 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 93,429 is an additive step under one definition; counter deltas against issue #13 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,298
OK in gate	9,298	-- of them directional	5,669
Out of gate (1w / 1M)	489 / 487	-- of them sideways	3,629
Invalid	16	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-11-09.json	Report cutoff	Mon Nov 16, 2026, 16:00 UTC
Generated at (pipeline)	2026-11-18 06:18 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,669 scored observations. MARKETMANIA RESEARCH TO DATE (as of Nov 16, 2026 cutoff): 93,429 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 77 published reports.

Issue #15. Weekly Model Watch is a living series. Issue #14's open questions -- "does grok-4.7 hold the top rank for a third issue, and does the title rule ever fire?" and "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?" -- read this issue as: gpt-6.1-sol at the top (51.1%), not the line that led issue #14 (grok-4.7), so the rank did not hold for a third issue; no weekly title was awarded, for the fifteenth issue running, and 3 of 4 qualifying turns found a caller, on turns in BTC / ETH; the two BTC turns fired outside ETH for a fifth issue running, and the second turn in each pair (Tue Nov 10, 08:00) drew callers again -- 14 on BTC and 29 on ETH -- while the first BTC turn (Mon Nov 9, 08:00) drew NOBODY. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the fourth 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 eleven 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 gpt-6.1-sol hold the top rank for a second issue, and does the title rule ever fire?
- Does a qualifying reversal fire outside ETH for a sixth issue running, and does a second turn in the same pair draw a caller for a third issue running?
- 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 #15	https://marketmania.ai/research/reports/consensus-watch-2026-11-09.pdf
Weekly Calibration #15	https://marketmania.ai/research/reports/weekly-calibration-2026-11-09.pdf
Weekly Model Watch #14	https://marketmania.ai/research/reports/model-watch-2026-11-02.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_2026w46,
  title  = {Weekly Model Watch #15: weekly leaderboard, ticker/pair reads and cross-confirmation, Nov 9-15 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {18},
  url    = {https://marketmania.ai/research/reports/model-watch-2026-11-09.pdf},
  note   = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Nov 9-15, 2026 UTC; source weekly_metrics_2026-11-09.json}
}
```