Weekly Model Watch #13

November 4, 2026 · MarketMania Research · Weekly series · window Oct 26-Nov 1, 2026 UTC

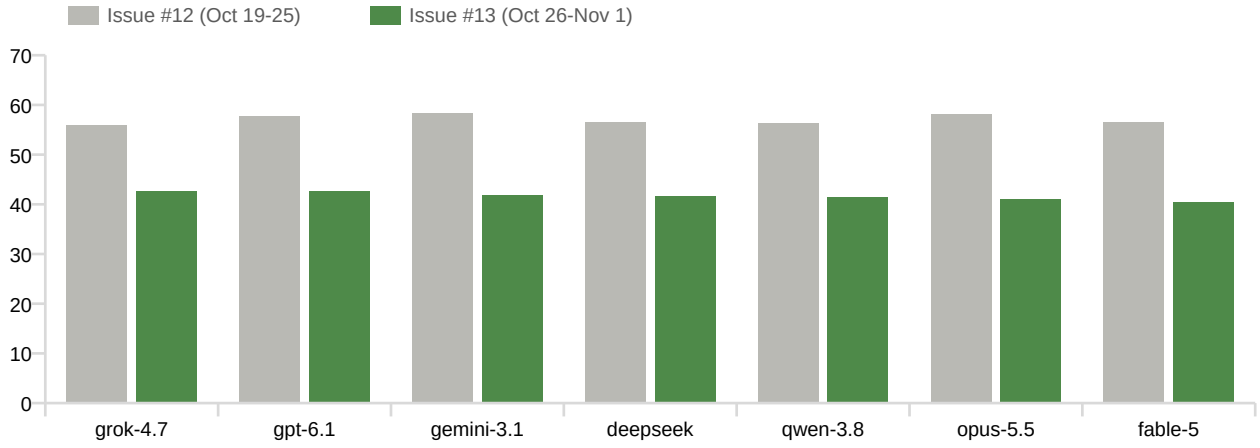
RESEARCH SNAPSHOT			
5,094 directional calls scored	**7** models (stable lineup)	of **9,304** mature forecasts	window **Oct 26-Nov 1, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **41.6%**	cutoff **Mon Nov 2, 2026, 16:00 UTC**
market: BTC net **-2.25%** (prior -2.39%)	ann vol **34.7%** (was 29.3%)	TOP5 volume **$14.86B**, +6% w/w	pairwise corr **0.93** (was 0.93)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 57.0% -> 41.6% (-15.3pp) with 0 of 7 models improving, and no weekly title was awarded, for the thirteenth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** grok-4.7 leads descriptively at 42.5% [39.2%, 45.9%]; the gap to #2 (gpt-6.1-sol, 42.5%) is 0.1pp. 0 of 7 models improved week over week; issue #12's leader gemini-3.1-pro is now 3 of 7.
- **No ticker cleared 51%.** SOL is the week's most readable coin (44.0%, n=1,099) and BNB the hardest at 36.8%; 0 of the 5 assets cleared 51%. Pair of the week: deepseek-v4-pro x SOL 46.5% (n=159); worst: claude-opus-5.5 x BNB 32.7% (n=110).
- **3 of 6 qualifying turns found a caller.** BTC short -2.36%, 0 caller(s), first NOBODY; BTC long +2.27%, 0 caller(s), first NOBODY; BTC short -4.55%, 14 caller(s), first gpt-6.1-sol; ETH short -3.46%, 2 caller(s), first gpt-6.1-sol; ETH long +2.36%, 0 caller(s), first NOBODY; ETH short -4.84%, 14 caller(s), first gemini-3.1-pro.
- **Self-agreement split across the four cells.** Cross-TF/FH lifts ran -4.0pp to +3.9pp (issue #12: +3.1pp to +9.2pp). On trend days, 4h TF-agreement hit 9.8% (n=112) against 38.1% (n=506) 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 57.0% -> 41.6%. "Was" values are the numbers published in issue #12; 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 #12, 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	830	62.4%	42.5%	39.2%-45.9%	-13.3
gpt-6.1-sol	829	62.3%	42.5%	39.1%-45.9%	-15.1
gemini-3.1-pro	755	56.8%	41.7%	38.3%-45.3%	-16.6
deepseek-v4-pro	745	56.0%	41.5%	38.0%-45.1%	-15.0
qwen-3.8-max	711	53.7%	41.3%	37.8%-45.0%	-15.0
claude-opus-5.5	591	44.4%	40.9%	37.0%-45.0%	-17.0
claude-fable-5	633	47.6%	40.4%	36.7%-44.3%	-15.9

No weekly title is awarded this issue -- for the thirteenth 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 0.1pp and their CIs overlap, so neither condition is met.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #12
SOL	1,099	44.0%	41.0%-46.9%	53.3%
XRP	909	42.7%	39.5%-45.9%	61.2%
BTC	1,136	42.3%	39.5%-45.2%	57.4%
ETH	1,032	41.8%	38.8%-44.8%	54.4%
BNB	918	36.8%	33.8%-40.0%	58.6%

SOL was the field's most readable ticker (44.0%) and BNB the hardest (36.8%); 0 of 5 assets cleared 51% this week (issue #12: 5 of 5). Issue #12's best ticker was XRP and its hardest SOL -- this issue's ordering is SOL > XRP > BTC > ETH > BNB.

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
deepseek-v4-pro	SOL	159	46.5%
gpt-6.1-sol	SOL	195	46.2%
claude-opus-5.5	SOL	125	45.6%
grok-4.7	XRP	155	44.5%
claude-opus-5.5	BTC	129	44.2%
grok-4.7	SOL	179	44.1%

Model (5 worst)	Symbol	N	Hit rate
claude-opus-5.5	ETH	130	39.2%
deepseek-v4-pro	BNB	134	34.3%
qwen-3.8-max	BNB	136	33.8%
claude-fable-5	BNB	108	33.3%
claude-opus-5.5	BNB	110	32.7%

deepseek-v4-pro x SOL (46.5%, n=159) is the week's best pair and claude-opus-5.5 x BNB (32.7%, n=110) the weakest. Pair history accumulates across issues -- read this as the thirteenth 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 Oct 28, 20:00	Short	-2.36%	0	NOBODY
BTC	Thu Oct 29, 08:00	Long	+2.27%	0	NOBODY
BTC	Thu Oct 29, 16:00	Short	-4.55%	14	gpt-6.1-sol
ETH	Wed Oct 28, 20:00	Short	-3.46%	2	gpt-6.1-sol
ETH	Thu Oct 29, 08:00	Long	+2.36%	0	NOBODY
ETH	Thu Oct 29, 16:00	Short	-4.84%	14	gemini-3.1-pro

The BTC short turn drew NOBODY; 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 Oct 30, 00:01 with confidence 70; the ETH short turn drew 2 caller(s), first gpt-6.1-sol on a 1d/1d call at Oct 29, 00:01 with confidence 61; the ETH long turn drew NOBODY; the ETH short turn drew 14 caller(s), first gemini-3.1-pro on a 1d/4h call at Oct 30, 00:01 with confidence 75. Scope in v1 remains BTC/ETH only; the event count across thirteen 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	618	33.0%	16	75.0%	37.0% (848)	-4.0
cross-TF	1d	TF 1d vs 4h	108	42.6%	6 *	0.0%	38.6% (163)	+3.9
cross-FH	4h	by FH 1h	588	34.4%	24	62.5%	37.0% (848)	-2.7
cross-FH	1d	by FH 4h	100	42.0%	6 *	0.0%	38.6% (163)	+3.3

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 214 / 49 / 235 / 57. Pooled across all four cell families the lifts run -4.0pp to +3.9pp this week (issue #12: +3.1pp to +9.2pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has trend on 1 of 7 days: 10-26 +0.10%, 10-27 +0.84%, 10-28 +0.60%, 10-29 -4.05% (trend), 10-30 -0.84%, 10-31 +0.21%, 11-01 +1.13%. On trend days, 4h TF-agreement hit 9.8% (n=112) against 38.1% (n=506) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	27.4%	20.2%	52.4%	1.36
claude-opus-5.5	1,330	21.7%	22.7%	55.6%	0.96
deepseek-v4-pro	1,330	21.9%	34.1%	44.0%	0.64
gemini-3.1-pro	1,330	22.8%	34.0%	43.2%	0.67
gpt-6.1-sol	1,330	18.6%	43.7%	37.7%	0.43
grok-4.7	1,330	23.8%	38.6%	37.6%	0.61
qwen-3.8-max	1,324	25.3%	28.4%	46.3%	0.89

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 0.79 this week. 'Sideways' was 45.2% of mature forecasts (4,210 of 9,304). Consensus skew: 611 symbol/FH/TF cells had 5 or more models on the same side and hit 40.2% (n=3,858) against the 41.6% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	BNB	Short	90	1h / 1h	Oct 30, 23:01	EXPIRY	-$0.65

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 grok-4.7 (42.5%); the lowest-Brier line is claude-opus-5.5 (Brier 0.2676, gap +15.4pp). Per-model ok-rate this window ran 99.59% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #13.

Market check: bigger daily moves on higher volume

OBSERVATION — mean |1d move| 1.35% -> 1.63% (+20% rel), BTC realized vol 27.5% -> 37.0% (ann., hourly); field directional accuracy 57.0% -> 41.6% (-15.3pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -$43.00 -> -$869.65 (adjacent calendar weeks Oct 19-25 vs Oct 26-Nov 1; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 56.0% -> 42.8% — the same direction as the trade-based hit rule.

Measure	Oct 19-25 (week A)	Oct 26-Nov 1 (week B)	Change
Mean |1d move| (5 assets)	1.35%	1.63%	+20% rel
BTC realized vol (ann., hourly)	27.5%	37.0%	+9.5pp
Field directional accuracy	57.0%	41.6%	-15.3pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	56.0%	42.8%	-13.2pp
Field sim win-rate	46.6%	33.6%	-13.0pp
Field sim net PnL	-$43.00	-$869.65	--

Week A is exactly the issue-#12 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 two 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 (2026-10-19..2026-10-25, the published issue-#12 window) and week B (2026-10-26..2026-11-01) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #12. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits nine 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. Thirteen 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): SOL leads this week at 44.0%, after two issues of XRP.
- Self-agreement lifts run -4.0pp to +3.9pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Oct 26-Nov 1, 2026 UTC). Descriptive for this window only; no claim about next week or any model's underlying skill. This window had trend on 1 of 7 days (issue #12: 1 of 7); every comparison with issue #12 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. 32 events across thirteen issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=16, 6, 24, 6); 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 3 still returns deepseek-v4-pro as the served model for our requests, with the same system fingerprint as on Oct 27, 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 8 this window). Series density and lineage notes (wave 13): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 26-Nov 1) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11 and #12: 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 488). (2) This window is the second 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 all 7 lines carry 1,470 rows each in both weeks, 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 two 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 two windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits nine 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 #12 after the audit that found the raw table mixes two exchanges; issue #12 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 2 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11 and #12, which printed 77,801 at the Oct 26 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 82,895 is an additive step under one definition; counter deltas against issue #11 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,304
OK in gate	9,304	-- of them directional	5,094
Out of gate (1w / 1M)	490 / 488	-- of them sideways	4,210
Invalid	8	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-26.json	Report cutoff	Mon Nov 2, 2026, 16:00 UTC
Generated at (pipeline)	2026-11-03 11:35 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,094 scored observations. MARKETMANIA RESEARCH TO DATE (as of Nov 2, 2026 cutoff): 82,895 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 65 published reports.

Issue #13. Weekly Model Watch is a living series. Issue #12's open questions -- "does gemini-3.1-pro 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 third issue running, and does a second turn in the same pair draw a caller again?" -- read this issue as: grok-4.7 at the top (42.5%), not the line that led issue #12 (gemini-3.1-pro), so the rank did not hold for a second issue; no weekly title was awarded, for the thirteenth issue running, and 3 of 6 qualifying turns found a caller, on turns in BTC / ETH; the three BTC turns fired outside ETH for a third issue running, and the second turn in each pair (Thu Oct 29, 08:00) drew NOBODY, while the third (Thu Oct 29, 16:00) drew 14 callers in each pair. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the second 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 nine 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 second issue, and does the title rule ever fire?
- Does a qualifying reversal fire outside ETH for a fourth issue running, and does a second turn in the same pair 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 #13	https://marketmania.ai/research/reports/consensus-watch-2026-10-26.pdf
Weekly Calibration #13	https://marketmania.ai/research/reports/weekly-calibration-2026-10-26.pdf
Weekly Model Watch #12	https://marketmania.ai/research/reports/model-watch-2026-10-19.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_2026w44,
  title  = {Weekly Model Watch #13: weekly leaderboard, ticker/pair reads and cross-confirmation, Oct 26-Nov 1 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {4},
  url    = {https://marketmania.ai/research/reports/model-watch-2026-10-26.pdf},
  note   = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Oct 26-Nov 1, 2026 UTC; source weekly_metrics_2026-10-26.json}
}
```