Weekly Model Watch #12

October 28, 2026 · MarketMania Research · Weekly series · window Oct 19-25, 2026 UTC

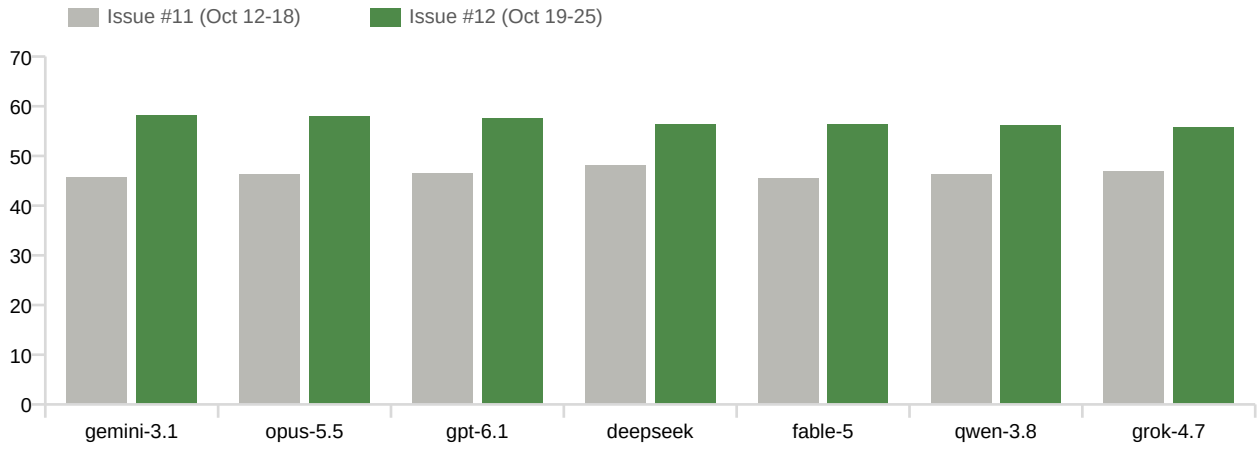
RESEARCH SNAPSHOT			
5,856 directional calls scored	**7** models (stable lineup)	of **9,171** mature forecasts	window **Oct 19-25, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **57.0%**	cutoff **Mon Oct 26, 2026, 16:00 UTC**
market: BTC net **-2.39%** (prior -1.98%)	ann vol **29.3%** (was 27.4%)	TOP5 volume **$14.02B**, -12% w/w	pairwise corr **0.93** (was 0.94)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 46.6% -> 57.0% (+10.4pp) with 7 of 7 models improving, and no weekly title was awarded, for the twelfth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** gemini-3.1-pro leads descriptively at 58.3% [55.0%, 61.5%]; the gap to #2 (claude-opus-5.5, 58.0%) is 0.3pp. 7 of 7 models improved week over week; issue #11's leader deepseek-v4-pro is now 4 of 7.
- **Every ticker was readable.** XRP is the week's most readable coin (61.2%, n=1,130) and SOL the hardest at 53.3%; 5 of the 5 assets cleared 51%. Pair of the week: gemini-3.1-pro x XRP 65.7% (n=166); worst: deepseek-v4-pro x ETH 50.3% (n=193).
- **3 of 3 qualifying turns found a caller.** BTC short -2.32%, 1 caller(s), first gpt-6.1-sol; ETH long +2.34%, 8 caller(s), first gemini-3.1-pro; ETH long +2.21%, 1 caller(s), first gemini-3.1-pro.
- **Self-agreement read as a lift in every cell.** Cross-TF/FH lifts ran +3.1pp to +9.2pp (issue #11: -6.2pp to -2.2pp). On trend days, 4h TF-agreement hit 74.8% (n=111) against 56.0% (n=591) 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 46.6% -> 57.0%. The opus-5.5, gpt-6.1 and grok-4.7 columns compare this window's single-build line against issue #11's spliced line. "Was" values are the numbers published in issue #11; 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. Issue #12's claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are single-build weeks while issue #11's were lineage splices, so read those three week-over-week rows as lines, not single model builds.

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. Week A (Oct 12-18, the issue-#11 window) is the spliced week: the three lines flipped at Oct 17, 2026 09:01 UTC inside it (gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5, grok-4.6 -> grok-4.7), so each week-over-week row for those lines joins this week's single-build line to last week's spliced line. This window (Oct 19-25) is the first full week on the new builds. In issue #11 those lines carried the labels "claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC)", "gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" and "grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)".

Leaderboard of the week

Model	N	Cov.	Hit rate	95% CI	W/w pp
gemini-3.1-pro	870	66.1%	58.3%	55.0%-61.5%	+12.5
claude-opus-5.5	700	52.6%	58.0%	54.3%-61.6%	+11.7
gpt-6.1-sol	962	72.3%	57.6%	54.4%-60.7%	+11.1
deepseek-v4-pro	891	67.0%	56.5%	53.2%-59.7%	+8.2
claude-fable-5	768	57.7%	56.4%	52.8%-59.9%	+10.8
qwen-3.8-max	829	62.4%	56.3%	52.9%-59.7%	+10.0
grok-4.7	836	69.4%	55.9%	52.5%-59.2%	+8.8

No weekly title is awarded this issue -- for the twelfth time. Title rule: **>=5pp gap AND non-overlapping 95% CI, N>=10** per side. gemini-3.1-pro's lead over claude-opus-5.5 is 0.3pp and their CIs overlap, so neither condition is met. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 week-over-week rows join this week's single-build line to issue #11's spliced line (the three lines flipped at Oct 17, 2026 09:01 UTC, inside week A) -- see How to read this.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #11
XRP	1,130	61.2%	58.3%-63.9%	49.0%
BNB	1,215	58.6%	55.8%-61.3%	44.8%
BTC	1,196	57.4%	54.6%-60.2%	44.0%
ETH	1,177	54.4%	51.5%-57.2%	47.4%
SOL	1,138	53.3%	50.4%-56.2%	47.4%

XRP was the field's most readable ticker (61.2%) and SOL the hardest (53.3%); 5 of 5 assets cleared 51% this week (issue #11: 0 of 5). Issue #11's best ticker was XRP and its hardest BTC -- this issue's ordering is XRP > BNB > BTC > ETH > 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
gemini-3.1-pro	XRP	166	65.7%
gpt-6.1-sol	XRP	187	62.6%
claude-opus-5.5	XRP	136	61.8%
claude-opus-5.5	BNB	151	60.9%
claude-fable-5	XRP	143	60.8%
deepseek-v4-pro	BTC	185	60.0%

Model (5 worst)	Symbol	N	Hit rate
claude-fable-5	SOL	158	51.9%
claude-opus-5.5	SOL	133	51.9%
grok-4.7	ETH	171	51.5%
gemini-3.1-pro	SOL	175	51.4%
deepseek-v4-pro	ETH	193	50.3%

gemini-3.1-pro x XRP (65.7%, n=166) is the week's best pair and deepseek-v4-pro x ETH (50.3%, n=193) the weakest. Pair history accumulates across issues -- read this as the twelfth 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	Thu Oct 22, 00:00	Short	-2.32%	1	gpt-6.1-sol
ETH	Tue Oct 20, 08:00	Long	+2.34%	8	gemini-3.1-pro
ETH	Sat Oct 24, 16:00	Long	+2.21%	1	gemini-3.1-pro

The BTC short turn drew 1 caller(s), first gpt-6.1-sol on a 4h/1h call at Oct 22, 00:01 with confidence 59; the ETH long turn drew 8 caller(s), first gemini-3.1-pro on a 1d/4h call at Oct 21, 00:01 with confidence 65; the ETH long turn drew 1 caller(s), first gemini-3.1-pro on a 4h/1h call at Oct 24, 16:01 with confidence 60. Scope in v1 remains BTC/ETH only; the event count across twelve 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	702	59.0%	34	41.2%	55.9% (923)	+3.1
cross-TF	1d	TF 1d vs 4h	95	75.8%	5 *	0.0%	67.5% (120)	+8.3
cross-FH	4h	by FH 1h	627	60.5%	44	40.9%	55.9% (923)	+4.5
cross-FH	1d	by FH 4h	73	76.7%	6 *	0.0%	67.5% (120)	+9.2

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 185 / 20 / 251 / 41. Pooled across all four cell families the lifts run +3.1pp to +9.2pp this week (issue #11: -6.2pp to -2.2pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has trend on 1 of 7 days: 10-19 -0.17%, 10-20 +1.24%, 10-21 +1.48%, 10-22 -3.34% (trend), 10-23 -0.27%, 10-24 -0.41%, 10-25 -0.91%. On trend days, 4h TF-agreement hit 74.8% (n=111) against 56.0% (n=591) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	35.5%	22.3%	42.3%	1.59
claude-opus-5.5	1,330	29.3%	23.3%	47.4%	1.26
deepseek-v4-pro	1,330	32.1%	34.9%	33.0%	0.92
gemini-3.1-pro	1,317	34.5%	31.6%	33.9%	1.09
gpt-6.1-sol	1,330	27.4%	44.9%	27.7%	0.61
grok-4.7	1,205	31.2%	38.2%	30.6%	0.82
qwen-3.8-max	1,329	34.1%	28.3%	37.6%	1.21

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 1.07 this week. 'Sideways' was 36.1% of mature forecasts (3,315 of 9,171). Consensus skew: 715 symbol/FH/TF cells had 5 or more models on the same side and hit 58.4% (n=4,569) against the 57.0% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	SOL	Long	85	1h / 1h	Oct 21, 22:01	EXPIRY	-$0.66

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 (58.3%); the lowest-Brier line is claude-opus-5.5 (Brier 0.2443, gap -1.5pp). Per-model ok-rate this window ran 90.00% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #12.

Market check: smaller daily moves on lower volume

OBSERVATION — mean |1d move| 2.05% -> 1.35% (-34% rel), BTC realized vol 34.9% -> 27.5% (ann., hourly); field directional accuracy 46.6% -> 57.0% (+10.4pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$352.09 -> -$43.00 (adjacent calendar weeks Oct 12-18 vs Oct 19-25; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 44.1% -> 56.0% — the same direction as the trade-based hit rule.

Measure	Oct 12-18 (week A)	Oct 19-25 (week B)	Change
Mean |1d move| (5 assets)	2.05%	1.35%	-34% rel
BTC realized vol (ann., hourly)	34.9%	27.5%	-7.4pp
Field directional accuracy	46.6%	57.0%	+10.4pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	44.1%	56.0%	+11.9pp
Field sim win-rate	39.2%	46.6%	+7.5pp
Field sim net PnL	-$352.09	-$43.00	--

Week A is exactly the issue-#11 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, inside week A, day 6 of 7) splices 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 (2026-10-12..2026-10-18, the published issue-#11 window) holds the cutover: its old-id rows are aggregated under the heads (it reproduces the published issue #11); week B (2026-10-19..2026-10-25) is the first full week on the heads (0 old-id row(s) remapped in it). The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits eight 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. Twelve 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): XRP led issue #11 and leads again at 61.2%.
- Self-agreement lifts run +3.1pp to +9.2pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Oct 19-25, 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 #11: 3 of 7); every comparison with issue #11 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. 26 events across twelve issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=34, 5, 44, 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 Oct 27 still returns deepseek-v4-pro as the served model for our requests, with the same system fingerprint as on Oct 21, 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 161 this window). Series density and lineage notes (wave 12): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 19-25) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10 and #11: 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 480, 1M 478). (2) This window is the first 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, and all 7 lines carry 1,470 rows each. Week A (the issue-#11 window) is the spliced week -- "claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the claude-opus-5 era, 315 from claude-opus-5.5); "gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the gpt-5.6-sol era, 315 from gpt-6.1-sol); "grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the grok-4.6 era, 315 from grok-4.7) -- so each of the three week-over-week rows joins this week's single-build line to last week's spliced line, and a remapped comparison is not the same model build. (3) No roll-out slot of the flip falls inside this window: the two hourly slots before it (Oct 17, 07:01 and 08:01 UTC) lie in week A, so this window's invalid counter holds none of their rows. (4) Past lineage flips: the Oct 17, 2026 09:01 UTC flip sits one window back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eight 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 #11 after the audit that found the raw table mixes two exchanges; issue #11 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 (Oct 26 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10 and #11, which printed 71,945 at the Oct 19 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 77,801 is an additive step under one definition; counter deltas against issue #10 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,171
OK in gate	9,171	-- of them directional	5,856
Out of gate (1w / 1M)	480 / 478	-- of them sideways	3,315
Invalid	161	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-19.json	Report cutoff	Mon Oct 26, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-27 11:37 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,856 scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 26, 2026 cutoff): 77,801 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 61 published reports.

Issue #12. Weekly Model Watch is a living series. Issue #11's open questions -- "does deepseek-v4-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 second issue running, and does a second turn in the same pair ever draw a caller?" -- read this issue as: gemini-3.1-pro at the top (58.3%), not the line that led issue #11 (deepseek-v4-pro), so the rank did not hold for a second issue; no weekly title was awarded, for the twelfth issue running, and 3 of 3 qualifying turns found a caller, on turns in BTC / ETH; the BTC turn (Thu Oct 22, 00:00) fired outside ETH for a second issue running, and the second of the two ETH turns (Sat Oct 24, 16:00) drew 1 caller. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the first full week on the Oct 17 builds; week-over-week rows compare against the spliced issue-#11 week, where those lines carried the labels "claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC)", "gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" and "grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)". The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eight 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 second issue, and does the title rule ever fire?
- 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?
- 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 #12	https://marketmania.ai/research/reports/consensus-watch-2026-10-19.pdf
Weekly Calibration #12	https://marketmania.ai/research/reports/weekly-calibration-2026-10-19.pdf
Weekly Model Watch #11	https://marketmania.ai/research/reports/model-watch-2026-10-12.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_2026w43,
  title  = {Weekly Model Watch #12: weekly leaderboard, ticker/pair reads and cross-confirmation, Oct 19-25 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {October}, day = {28},
  url    = {https://marketmania.ai/research/reports/model-watch-2026-10-19.pdf},
  note   = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Oct 19-25, 2026 UTC; source weekly_metrics_2026-10-19.json}
}
```