

Weekly Model Watch #10

October 14, 2026 · MarketMania Research · Weekly series · window Oct 5-11, 2026 UTC

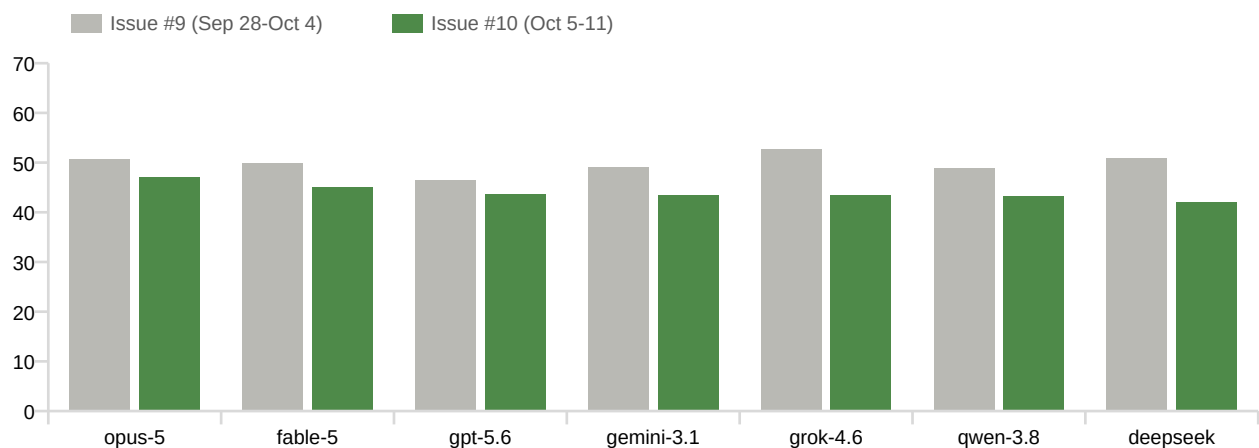
RESEARCH SNAPSHOT			
4,601 directional calls scored	7 models (stable lineup)	of 9,249 mature forecasts	window Oct 5-11, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 43.8%	cutoff Mon Oct 12, 2026, 16:00 UTC
market: BTC net -3.19% (prior +2.44%)	ann vol 27.2% (was 21.1%)	TOP5 volume \$15.97B , -11% w/w	pairwise corr 0.85 (was 0.57)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 49.6% -> 43.8% (-5.9pp) with 0 of 7 models improving, and no weekly title was awarded, for the tenth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** claude-opus-5 leads descriptively at 47.0% [42.7%, 51.3%]; the gap to #2 (claude-fable-5, 45.0%) is 2.0pp. 0 of 7 models improved week over week; issue #9's leader grok-4.6 is now 5 of 7.
- **No ticker cleared 51%.** BTC is the week's most readable coin (46.1%, n=975) and XRP the hardest at 42.1%; 0 of the 5 assets cleared 51%. Pair of the week: claude-opus-5 x XRP 49.6% (n=117); worst: deepseek-v4-pro x XRP 37.8% (n=156).
- **No qualifying reversal fired.** No BTC/ETH turn met the detector threshold inside this window (detector status no_reversal_events), so there is nothing to score this issue.
- **Self-agreement split across the four cells.** Cross-TF/FH lifts ran -2.0pp to +6.8pp (issue #9: +0.4pp to +18.7pp). On trend days, 4h TF-agreement hit 66.7% (n=237) against 26.2% (n=302) 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 49.6% -> 43.8%. The grok column compares two pure grok-4.6 weeks: the 4.5 -> 4.6 flip sits six windows back. "Was" values are the numbers published in issue #9; 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 #9 as well -- the 4.5 -> 4.6 flip sits six 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 six 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 fifth 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
claude-opus-5	517	39.8%	47.0%	42.7%-51.3%	-3.7
claude-fable-5	613	47.1%	45.0%	41.1%-49.0%	-4.9
gpt-5.6-sol	734	55.2%	43.5%	39.9%-47.1%	-3.1
gemini-3.1-pro	722	54.3%	43.4%	39.8%-47.0%	-5.6
grok-4.6	547	41.1%	43.3%	39.2%-47.5%	-9.3
qwen-3.8-max	707	53.2%	43.1%	39.5%-46.8%	-5.8
deepseek-v4-pro	761	57.2%	42.0%	38.6%-45.6%	-8.9

No weekly title is awarded this issue -- for the tenth time. Title rule: ≥ 5 pp gap AND non-overlapping 95% CI, $N \geq 10$ per side. claude-opus-5's lead over claude-fable-5 is 2.0pp and their CIs overlap, so neither condition is met. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits six 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 fifth pure-4.6 against pure-4.6 comparison of the series.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #9
BTC	975	46.1%	42.9%-49.2%	53.3%
SOL	895	44.2%	41.0%-47.5%	46.0%
ETH	896	43.9%	40.6%-47.1%	50.8%
BNB	877	42.4%	39.2%-45.7%	47.9%
XRP	958	42.1%	39.0%-45.2%	50.2%

BTC was the field's most readable ticker (46.1%) and XRP the hardest (42.1%); 0 of 5 assets cleared 51% this week (issue #9: 1 of 5). Issue #9's best ticker was BTC and its hardest SOL -- this issue's ordering is BTC > SOL > ETH > BNB > XRP.

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
claude-opus-5	XRP	117	49.6%
claude-fable-5	ETH	122	49.2%
claude-opus-5	SOL	95	48.4%
gemini-3.1-pro	BTC	154	48.0%
claude-opus-5	ETH	100	48.0%
grok-4.6	BTC	107	47.7%

Model (5 worst)	Symbol	N	Hit rate
qwen-3.8-max	BNB	134	40.3%
gpt-5.6-sol	XRP	149	40.3%
grok-4.6	BNB	102	40.2%
gemini-3.1-pro	XRP	142	38.0%
deepseek-v4-pro	XRP	156	37.8%

claude-opus-5 x XRP (49.6%, n=117) is the week's best pair and deepseek-v4-pro x XRP (37.8%, n=156) the weakest. Pair history accumulates across issues -- read this as the tenth 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].

No qualifying reversal fired inside this window (status: no_reversal_events), so the block is empty by construction, not by failure to look.

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	539	44.0%	40	62.5%	45.8% (797)	-1.8
cross-TF	1d	TF 1d vs 4h	111	24.3%	2 *	0.0%	21.2% (160)	+3.1
cross-FH	4h	by FH 1h	493	43.8%	38	57.9%	45.8% (797)	-2.0
cross-FH	1d	by FH 4h	89	28.1%	2 *	0.0%	21.2% (160)	+6.8

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 218 / 47 / 266 / 69. Pooled across all four cell families the lifts run -2.0pp to +6.8pp this week (issue #9: +0.4pp to +18.7pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has 2 trend days of 7: 10-05 -0.61%, 10-06 -0.37%, 10-07 -2.65% (trend), 10-08 -1.81% (trend), 10-09 +1.06%, 10-10 +0.48%, 10-11 +0.93%. On trend days, 4h TF-agreement hit 66.7% (n=237) against 26.2% (n=302) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,300	27.1%	20.1%	52.8%	1.35
claude-opus-5	1,300	20.5%	19.3%	60.2%	1.06
deepseek-v4-pro	1,330	20.8%	36.5%	42.8%	0.57
gemini-3.1-pro	1,330	27.9%	26.4%	45.7%	1.06
gpt-5.6-sol	1,330	20.4%	34.8%	44.8%	0.59
grok-4.6	1,330	12.9%	28.2%	58.9%	0.46
qwen-3.8-max	1,329	24.5%	28.7%	46.8%	0.86

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 0.85 this week. 'Sideways' was 50.3% of mature forecasts (4,648 of 9,249). Consensus skew: 508 symbol/FH/TF cells had 5 or more models on the same side and hit 43.6% (n=3,171) against the 43.8% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gpt-5.6-sol	BTC	Short	82	4h / 1h	Oct 8, 16:01	EXPIRY	-\$0.97

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

claude-opus-5 is both this week's hit-rate leader (47.0%) and the lowest-Brier line (Brier 0.2685, gap +13.8pp). Per-model ok-rate this window ran 97.96% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #10.

Market check: a down week with bigger daily moves

OBSERVATION — mean |1d move| 0.87% -> 1.54% (+77% rel), BTC realized vol 32.2% -> 30.9% (ann., hourly); field directional accuracy 49.6% -> 43.8% (-5.9pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -\$408.24 -> -\$520.82 (adjacent calendar weeks Sep 28-Oct 4 vs Oct 5-11; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 49.0% -> 45.5% — the same direction as the trade-based hit rule.

Measure	Sep 28-Oct 4 (week A)	Oct 5-11 (week B)	Change
Mean 1d move (5 assets)	0.87%	1.54%	+77% rel
BTC realized vol (ann., hourly)	32.2%	30.9%	-1.3pp
Field directional accuracy	49.6%	43.8%	-5.9pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	49.0%	45.5%	-3.5pp
Field sim win-rate	41.0%	34.5%	-6.5pp
Field sim net PnL	-\$408.24	-\$520.82	--

Week A is exactly the issue-#9 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, six windows back, inside issue #4's window). Week A (2026-09-28..2026-10-04) is pure grok-4.6 (it reproduces the published issue #9); 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. Ten issues, and the best ticker changed hands in every one except issues #6 and #10 (SOL -> BNB -> BTC -> SOL -> BNB -> BNB -> ETH -> XRP -> BTC -> BTC): BTC led issue #9 and leads again at 46.1%.
- Self-agreement lifts run -2.0pp to +6.8pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Oct 5-11, 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 #9: 2 of 7); every comparison with issue #9 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. 17 events across ten issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=40, 2, 38, 2); 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. DeepSeek announced routing of V4 Pro to V4.1 Flash from Sep 14, 04:00 UTC; an API check on Oct 13 still returns deepseek-v4-pro as the served model for our requests, with the same system fingerprint as on Oct 7, 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 72 this window). Series density and lineage notes (wave 10): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 5-11) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8 and #9: 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 479). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, six 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-#9 window), so every grok number in this issue and in issue #9 is a pure grok-4.6 line -- the grok week-over-week row is the fifth 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 #9 after the audit that found the raw table mixes two exchanges; issue #9 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 12 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8 and #9, which printed 62,260 at the Oct 5 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 66,861 is an additive step under one definition; counter deltas against issue #8 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,249
OK in gate	9,249	-- of them directional	4,601
Out of gate (1w / 1M)	490 / 479	-- of them sideways	4,648
Invalid	72	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-05.json	Report cutoff	Mon Oct 12, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-13 15:15 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,601 scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 12, 2026 cutoff): 66,861 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 53 published reports.

Issue #10. Weekly Model Watch is a living series. Issue #9's open questions -- "does grok-4.6 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 ever draw a caller?" -- read this issue as: claude-opus-5 at the top (47.0%), not the line that led issue #9 (grok-4.6), so the rank did not hold for a second issue; no weekly title was awarded, for the tenth issue running, and no qualifying reversal fired, so the run of issues with a qualifying turn outside ETH stops at two (issues #8 and #9) and the second-turn question stays open. The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #9; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits six 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 claude-opus-5 hold the top rank for a second issue, and does the title rule ever fire?
- Does a qualifying reversal fire again after an issue with none, 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 #10	https://marketmania.ai/research/reports/consensus-watch-2026-10-05.pdf
Weekly Calibration #10	https://marketmania.ai/research/reports/weekly-calibration-2026-10-05.pdf
Weekly Model Watch #9	https://marketmania.ai/research/reports/model-watch-2026-09-28.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_2026w41,
  title = {Weekly Model Watch #10: weekly leaderboard, ticker/pair reads and cross-confirmation, Oct 5-11 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {October}, day = {14},
  url   = {https://marketmania.ai/research/reports/model-watch-2026-10-05.pdf},
  note  = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Oct 5-11, 2026 UTC; source weekly_metrics_2026-10-05.json}
}
```