

Weekly Model Watch #9

October 7, 2026 · MarketMania Research · Weekly series · window Sep 28-Oct 4, 2026 UTC

RESEARCH SNAPSHOT

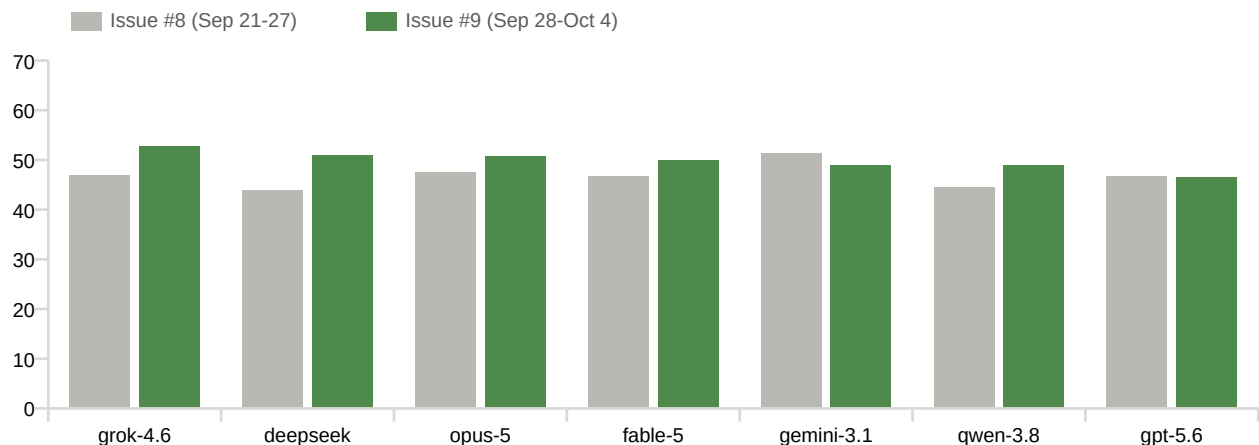
4,331 directional calls scored	7 models (stable lineup)	of 9,286 mature forecasts	window Sep 28-Oct 4, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 49.6%	cutoff Mon Oct 5, 2026, 16:00 UTC
market: BTC net +2.44% (prior +4.06%)	ann vol 21.1% (was 52.2%)	TOP5 volume \$17.96B , -20% w/w	pairwise corr 0.57 (was 0.71)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 46.8% -> 49.6% (+2.8pp) with 5 of 7 models improving, and no weekly title was awarded, for the ninth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** grok-4.6 leads descriptively at 52.7% [47.9%, 57.3%]; the gap to #2 (deepseek-v4-pro, 50.9%) is 1.7pp. 5 of 7 models improved week over week; issue #8's leader gemini-3.1-pro is now 5 of 7.
- **1 of 5 tickers were readable.** BTC is the week's most readable coin (53.3%, n=944) and SOL the hardest at 46.0%; 1 of the 5 assets cleared 51%. Pair of the week: grok-4.6 x BTC 58.1% (n=93); worst: gpt-5.6-sol x SOL 41.5% (n=130).
- **0 of 2 qualifying turns found a caller.** BTC short -2.28%, 0 caller(s), first NOBODY; ETH short -2.75%, 0 caller(s), first NOBODY.
- **Self-agreement read as a lift in every cell.** Cross-TF/FH lifts ran +0.4pp to +18.7pp (issue #8: -2.5pp to +3.8pp). On trend days, 4h TF-agreement hit 66.9% (n=121) against 42.4% (n=321) 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 46.8% -> 49.6%. The grok column compares two pure grok-4.6 weeks: the 4.5 -> 4.6 flip sits five windows back. "Was" values are the numbers published in issue #8; 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 #8 as well -- the 4.5 -> 4.6 flip sits five 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 five 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 fourth 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
grok-4.6	431	32.6%	52.7%	47.9%-57.3%	+5.7
deepseek-v4-pro	687	51.6%	50.9%	47.2%-54.7%	+7.1
claude-opus-5	534	40.2%	50.7%	46.5%-55.0%	+3.2
claude-fable-5	623	46.8%	49.9%	46.0%-53.8%	+3.2
gemini-3.1-pro	694	52.2%	49.0%	45.3%-52.7%	-2.4
qwen-3.8-max	711	53.5%	48.9%	45.3%-52.6%	+4.4
gpt-5.6-sol	651	49.5%	46.5%	42.7%-50.4%	-0.2

No weekly title is awarded this issue -- for the ninth time. Title rule: ≥ 5 pp gap AND non-overlapping 95% CI, $N \geq 10$ per side. grok-4.6's lead over deepseek-v4-pro is 1.7pp and their CIs overlap, so neither condition is met. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits five 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 fourth pure-4.6 against pure-4.6 comparison of the series.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #8
BTC	944	53.3%	50.1%-56.5%	48.9%
ETH	786	50.8%	47.3%-54.2%	45.8%
XRP	806	50.2%	46.8%-53.7%	50.0%
BNB	899	47.9%	44.7%-51.2%	38.2%
SOL	896	46.0%	42.7%-49.3%	48.7%

BTC was the field's most readable ticker (53.3%) and SOL the hardest (46.0%); 1 of 5 assets cleared 51% this week (issue #8: 0 of 5). Issue #8's best ticker was XRP and its hardest BNB -- this issue's ordering is BTC > ETH > XRP > 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
grok-4.6	BTC	93	58.1%
qwen-3.8-max	BTC	145	56.5%
claude-fable-5	BTC	131	55.7%
grok-4.6	ETH	73	54.8%
grok-4.6	XRP	81	54.3%
deepseek-v4-pro	XRP	125	53.6%

Model (5 worst)	Symbol	N	Hit rate
qwen-3.8-max	BNB	150	46.7%
gpt-5.6-sol	BNB	136	44.9%
qwen-3.8-max	SOL	144	44.4%
claude-fable-5	SOL	135	43.0%
gpt-5.6-sol	SOL	130	41.5%

grok-4.6 x BTC (58.1%, n=93) is the week's best pair and gpt-5.6-sol x SOL (41.5%, n=130) the weakest. Pair history accumulates across issues -- read this as the ninth 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	Fri Oct 2, 12:00	Short	-2.28%	0	NOBODY
ETH	Fri Oct 2, 12:00	Short	-2.75%	0	NOBODY

The BTC short turn drew NOBODY; the ETH short turn drew NOBODY. Scope in v1 remains BTC/ETH only; the event count across nine 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	442	49.1%	5 *	0.0%	48.7% (548)	+0.4
cross-TF	1d	TF 1d vs 4h	33	63.6%	1 *	100.0%	57.3% (96)	+6.3
cross-FH	4h	by FH 1h	404	53.2%	14	21.4%	48.7% (548)	+4.5
cross-FH	1d	by FH 4h	25	76.0%	1 *	100.0%	57.3% (96)	+18.7

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 100 / 62 / 129 / 70. Pooled across all four cell families the lifts run +0.4pp to +18.7pp this week (issue #8: -2.5pp to +3.8pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has 2 trend days of 7: 09-28 -1.19%, 09-29 +0.22%, 09-30 -0.22%, 10-01 +1.54% (trend), 10-02 -0.39%, 10-03 +0.40%, 10-04 +1.81% (trend). On trend days, 4h TF-agreement hit 66.9% (n=121) against 42.4% (n=321) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	32.0%	14.8%	53.2%	2.16
claude-opus-5	1,330	27.4%	12.8%	59.9%	2.14
deepseek-v4-pro	1,330	31.1%	20.5%	48.4%	1.52
gemini-3.1-pro	1,330	31.9%	20.2%	47.8%	1.58
gpt-5.6-sol	1,315	28.4%	21.1%	50.5%	1.34
grok-4.6	1,322	19.1%	13.5%	67.4%	1.41
qwen-3.8-max	1,329	34.4%	19.1%	46.5%	1.80

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 1.71 this week. 'Sideways' was 53.4% of mature forecasts (4,955 of 9,286). Consensus skew: 501 symbol/FH/TF cells had 5 or more models on the same side and hit 51.1% (n=3,247) against the 49.6% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	XRP	Long	85	1h / 1h	Oct 4, 08:01	EXPIRY	-\$0.29

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

grok-4.6 is both this week's hit-rate leader (52.7%) and the lowest-Brier line (Brier 0.2566, gap +7.7pp). Per-model ok-rate this window ran 98.98% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #9.

Market check: an up week on lower volatility

OBSERVATION — mean |1d move| 1.83% -> 0.87% (-52% rel), BTC realized vol 35.8% -> 32.2% (ann., hourly); field directional accuracy 46.8% -> 49.6% (+2.8pp), 5 of 7 models improved, sim win-rate up for 5 of 7, field sim PnL -\$315.68 -> -\$408.24 (adjacent calendar weeks Sep 21-27 vs Sep 28-Oct 4; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 49.2% -> 49.0% — the opposite direction to the trade-based hit rule.

Measure	Sep 21-27 (week A)	Sep 28-Oct 4 (week B)	Change
Mean 1d move (5 assets)	1.83%	0.87%	-52% rel
BTC realized vol (ann., hourly)	35.8%	32.2%	-3.6pp
Field directional accuracy	46.8%	49.6%	+2.8pp
Models improving hit-rate	--	5 of 7	--
Raw price-sign accuracy	49.2%	49.0%	-0.2pp
Field sim win-rate	40.0%	41.0%	+1.1pp
Field sim net PnL	-\$315.68	-\$408.24	--

Week A is exactly the issue-#8 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, five windows back, inside issue #4's window). Week A (2026-09-21..2026-09-27) is pure grok-4.6 (it reproduces the published issue #8); 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. Nine issues, and the best ticker changed hands in every one except issue #6 (SOL -> BNB -> BTC -> SOL -> BNB -> BNB -> ETH -> XRP -> BTC): BTC leads this week at 53.3%, after one issue of XRP.
- Self-agreement lifts run +0.4pp to +18.7pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Sep 28-Oct 4, 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 #8: 2 of 7); every comparison with issue #8 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 nine issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=5, 1, 14, 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, 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 7 still returns deepseek-v4-pro as the served model for our requests, with the same system fingerprint as on Sep 29, so this issue keeps the label; it will change the day the served id changes. On Sep 21, from 04:01 UTC, the provider returned HTTP 402 (insufficient balance) to 20 deepseek-v4-pro calls inside week A (the issue-#8 window), counted as invalid there (invalid 44 in issue #8); in this window the line's ok-rate is 100.0% (invalid 28 this window). Series density and lineage notes (wave 9): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 28-Oct 4) daily coverage is FULL, as it was in issues #4, #5, #6, #7 and #8: 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 486). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, five 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-#8 window), so every grok number in this issue and in issue #8 is a pure grok-4.6 line -- the grok week-over-week row is the fourth 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 #8 after the audit that found the raw table mixes two exchanges; issue #8 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 5 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7 and #8, which printed 57,929 at the Sep 28 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 62,260 is an additive step under one definition; counter deltas against issue #7 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,286
OK in gate	9,286	-- of them directional	4,331
Out of gate (1w / 1M)	490 / 486	-- of them sideways	4,955
Invalid	28	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-28.json	Report cutoff	Mon Oct 5, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-07 06:16 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,331 scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 5, 2026 cutoff): 62,260 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 45 published reports.

Issue #9. Weekly Model Watch is a living series. Issue #8's open questions -- "does gemini-3.1-pro hold the top rank for a fifth 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: grok-4.6 at the top (52.7%), not the line that led issues #5 to #8 (gemini-3.1-pro), so the rank did not hold for a fifth issue; no weekly title was awarded, for the ninth issue running, and 0 of 2 qualifying turns found a caller, on turns in BTC / ETH; the BTC turn makes a second issue running with a qualifying turn outside ETH, and with one turn per pair the second-turn question stays open. The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #8; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits five 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-4.6 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 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 #9	https://marketmania.ai/research/reports/consensus-watch-2026-09-28.pdf
Weekly Calibration #9	https://marketmania.ai/research/reports/weekly-calibration-2026-09-28.pdf
Weekly Model Watch #8	https://marketmania.ai/research/reports/model-watch-2026-09-21.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_2026w40,
  title = {Weekly Model Watch #9: weekly leaderboard, ticker/pair reads and crossconfirmation, Sep 28-Oct 4 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {October}, day = {7},
  url   = {https://marketmania.ai/research/reports/model-watch-2026-09-28.pdf},
  note  = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Sep 28-Oct 4, 2026 UTC; source weekly_metrics_2026-09-28.json}
}
```