

Weekly Model Watch #8

September 30, 2026 · MarketMania Research · Weekly series · window Sep 21-27, 2026 UTC

RESEARCH SNAPSHOT

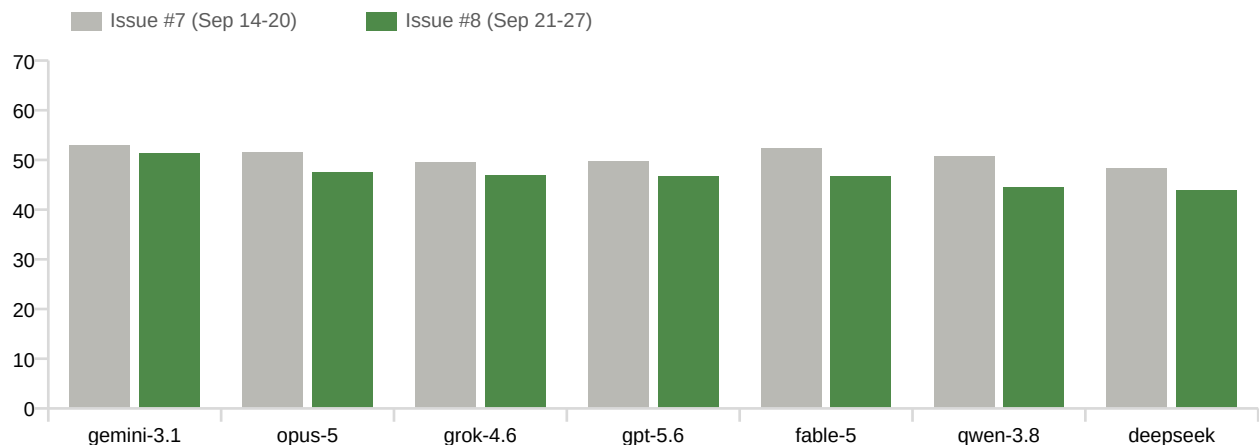
4,958 directional calls scored	7 models (stable lineup)	of 9,273 mature forecasts	window Sep 21-27, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 46.8%	cutoff Mon Sep 28, 2026, 16:00 UTC
market: BTC net +4.06% (prior +5.64%)	ann vol 52.2% (was 51.0%)	TOP5 volume \$22.43B , +20% w/w	pairwise corr 0.71 (was 0.93)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 50.7% -> 46.8% (-3.9pp) with 0 of 7 models improving, and no weekly title was awarded, for the eighth issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** gemini-3.1-pro leads descriptively at 51.4% [47.9%, 54.9%]; the gap to #2 (claude-opus-5, 47.6%) is 3.8pp. 0 of 7 models improved week over week; issue #7's leader gemini-3.1-pro is now 1 of 7.
- **No ticker cleared 51%.** XRP is the week's most readable coin (50.0%, n=1,165) and BNB the hardest at 38.2%; 0 of the 5 assets cleared 51%. Pair of the week: gemini-3.1-pro x SOL 54.3% (n=186); worst: claude-fable-5 x BNB 36.0% (n=114).
- **2 of 2 qualifying turns found a caller.** BTC short -2.90%, 1 caller(s), first gemini-3.1-pro; ETH short -3.73%, 6 caller(s), first gpt-5.6-sol.
- **Self-agreement split across the four cells.** Cross-TF/FH lifts ran -2.5pp to +3.8pp (issue #7: +1.4pp to +6.5pp). On trend days, 4h TF-agreement hit 64.0% (n=264) against 27.7% (n=314) 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 50.7% -> 46.8%. The grok column compares two pure grok-4.6 weeks: the 4.5 -> 4.6 flip sits four windows back. "Was" values are the numbers published in issue #7; 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 #7 as well -- the 4.5 -> 4.6 flip sits four 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 four 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 third 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
gemini-3.1-pro	794	59.7%	51.4%	47.9%-54.9%	-1.5
claude-opus-5	620	46.6%	47.6%	43.7%-51.5%	-4.0
grok-4.6	537	40.6%	46.9%	42.7%-51.2%	-2.7
gpt-5.6-sol	744	55.9%	46.8%	43.2%-50.4%	-2.9
claude-fable-5	736	55.3%	46.7%	43.2%-50.3%	-5.6
qwen-3.8-max	750	56.8%	44.5%	41.0%-48.1%	-6.3
deepseek-v4-pro	777	59.3%	43.9%	40.4%-47.4%	-4.5

No weekly title is awarded this issue -- for the eighth time. Title rule: ≥ 5 pp gap AND non-overlapping 95% CI, $N \geq 10$ per side. gemini-3.1-pro's lead over claude-opus-5 is 3.8pp and their CIs overlap, so neither condition is met. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits four 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 third pure-4.6 against pure-4.6 comparison of the series.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #7
XRP	1,165	50.0%	47.1%-52.8%	50.4%
BTC	919	48.9%	45.6%-52.1%	50.3%
SOL	1,190	48.7%	45.9%-51.6%	49.5%
ETH	891	45.8%	42.5%-49.1%	57.2%
BNB	793	38.2%	34.9%-41.6%	46.3%

XRP was the field's most readable ticker (50.0%) and BNB the hardest (38.2%); 0 of 5 assets cleared 51% this week (issue #7: 1 of 5). Issue #7's best ticker was ETH and its hardest BNB -- this issue's ordering is XRP > BTC > SOL > 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
gemini-3.1-pro	SOL	186	54.3%
gemini-3.1-pro	BTC	154	53.2%
gemini-3.1-pro	XRP	181	53.0%
gemini-3.1-pro	ETH	142	52.1%
gpt-5.6-sol	XRP	177	52.0%
grok-4.6	XRP	136	51.5%

Model (5 worst)	Symbol	N	Hit rate
grok-4.6	BNB	86	38.4%
deepseek-v4-pro	BNB	124	37.1%
qwen-3.8-max	BNB	130	36.9%
claude-opus-5	BNB	86	36.0%
claude-fable-5	BNB	114	36.0%

gemini-3.1-pro x SOL (54.3%, n=186) is the week's best pair and claude-fable-5 x BNB (36.0%, n=114) the weakest. Pair history accumulates across issues -- read this as the eighth 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 Sep 23, 08:00	Short	-2.90%	1	gemini-3.1-pro
ETH	Wed Sep 23, 08:00	Short	-3.73%	6	gpt-5.6-sol

The BTC short turn drew 1 caller(s), first gemini-3.1-pro on a 1d/4h call at Sep 24, 00:01 with confidence 65; the ETH short turn drew 6 caller(s), first gpt-5.6-sol on a 4h/1h call at Sep 23, 08:01 with confidence 67. Scope in v1 remains BTC/ETH only; the event count across eight 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	578	44.3%	21	42.9%	46.8% (773)	-2.5
cross-TF	1d	TF 1d vs 4h	135	48.9%	8 *	50.0%	48.1% (206)	+0.8
cross-FH	4h	by FH 1h	513	45.8%	25	40.0%	46.8% (773)	-1.0
cross-FH	1d	by FH 4h	106	51.9%	9 *	55.6%	48.1% (206)	+3.8

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 173 / 63 / 234 / 91. Pooled across all four cell families the lifts run -2.5pp to +3.8pp this week (issue #7: +1.4pp to +6.5pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has 2 trend days of 7: 09-21 +6.39% (trend), 09-22 -0.40%, 09-23 -2.04% (trend), 09-24 +0.02%, 09-25 -0.46%, 09-26 +0.36%, 09-27 +0.20%. On trend days, 4h TF-agreement hit 64.0% (n=264) against 27.7% (n=314) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	43.8%	11.6%	44.7%	3.78
claude-opus-5	1,330	36.5%	10.2%	53.4%	3.59
deepseek-v4-pro	1,310	41.5%	17.8%	40.7%	2.33
gemini-3.1-pro	1,330	41.1%	18.6%	40.3%	2.21
gpt-5.6-sol	1,330	37.1%	18.8%	44.1%	1.98
grok-4.6	1,323	26.3%	14.3%	59.4%	1.84
qwen-3.8-max	1,320	41.0%	15.8%	43.2%	2.59

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 2.62 this week. 'Sideways' was 46.5% of mature forecasts (4,315 of 9,273). Consensus skew: 606 symbol/FH/TF cells had 5 or more models on the same side and hit 46.5% (n=3,902) against the 46.8% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gemini-3.1-pro	BTC	Long	80	1d / 1d	Sep 23, 00:01	SL	-\$3.10

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 gemini-3.1-pro (51.4%); the lowest-Brier line is qwen-3.8-max (Brier 0.2689, gap +15.6pp). Per-model ok-rate this window ran 98.64% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #8.

Market check: an up week with smaller daily moves

OBSERVATION — mean |1d move| 2.69% -> 1.83% (-32% rel), BTC realized vol 35.1% -> 35.8% (ann., hourly); field directional accuracy 50.7% -> 46.8% (-3.9pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -\$337.72 -> -\$315.68 (adjacent calendar weeks Sep 14-20 vs Sep 21-27; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 54.2% -> 49.2% — the same direction as the trade-based hit rule.

Measure	Sep 14-20 (week A)	Sep 21-27 (week B)	Change
Mean 1d move (5 assets)	2.69%	1.83%	-32% rel
BTC realized vol (ann., hourly)	35.1%	35.8%	+0.7pp
Field directional accuracy	50.7%	46.8%	-3.9pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	54.2%	49.2%	-4.9pp
Field sim win-rate	43.4%	40.0%	-3.4pp
Field sim net PnL	-\$337.72	-\$315.68	--

Week A is exactly the issue-#7 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, four windows back, inside issue #4's window). Week A (2026-09-14..2026-09-20) is pure grok-4.6 (it reproduces the published issue #7); 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. Eight issues, and the best ticker changed hands in every one except issue #6 (SOL -> BNB -> BTC -> SOL -> BNB -> BNB -> ETH -> XRP): XRP leads this week at 50.0%, after one issue of ETH.
- Self-agreement lifts run -2.5pp to +3.8pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Sep 21-27, 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 #7: 3 of 7); every comparison with issue #7 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. 15 events across eight issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=21, 8, 25, 9); 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 Sep 29 still returns deepseek-v4-pro as the served model for our requests, with the same system fingerprint as on Sep 22, 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 this window; they are counted as invalid (invalid 44 this window; the line's ok-rate 98.6%). Series density and lineage notes (wave 8): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 21-27) daily coverage is FULL, as it was in issues #4, #5, #6 and #7: 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 488, 1M 485). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, four 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-#7 window), so every grok number in this issue and in issue #7 is a pure grok-4.6 line -- the grok week-over-week row is the third 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 #7 after the audit that found the raw table mixes two exchanges; issue #7 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 (Sep 28 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6 and #7, which printed 52,971 at the Sep 21 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 57,929 is an additive step under one definition; counter deltas against issue #6 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,273
OK in gate	9,273	-- of them directional	4,958
Out of gate (1w / 1M)	488 / 485	-- of them sideways	4,315
Invalid	44	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-21.json	Report cutoff	Mon Sep 28, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-29 08:16 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,958 scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 28, 2026 cutoff): 57,929 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 37 published reports.

Issue #8. Weekly Model Watch is a living series. Issue #7's open questions -- "does gemini-3.1-pro hold the top rank for a fourth issue, and does the title rule ever fire?" and "Does a qualifying reversal ever fire outside ETH, and does a second turn in the same pair ever draw a caller?" -- read this issue as: gemini-3.1-pro at the top (51.4%), the line that also led issue #7 (gemini-3.1-pro), so the rank held for a fourth issue; no weekly title was awarded, for the eighth issue running, and 2 of 2 qualifying turns found a caller, on turns in BTC / ETH; the BTC turn is the first qualifying turn outside ETH since issue #4, 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 #7; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits four 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 fifth issue, and does the title rule ever fire?
- 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?
- 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 #8	https://marketmania.ai/research/reports/consensus-watch-2026-09-21.pdf
Weekly Calibration #8	https://marketmania.ai/research/reports/weekly-calibration-2026-09-21.pdf
Weekly Model Watch #7	https://marketmania.ai/research/reports/model-watch-2026-09-14.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_2026w39,
  title = {Weekly Model Watch #8: weekly leaderboard, ticker/pair reads and cross-confirmation, Sep 21-27 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {30},
  url = {https://marketmania.ai/research/reports/model-watch-2026-09-21.pdf},
  note = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Sep 21-27, 2026 UTC; source weekly_metrics_2026-09-21.json}
}
```