

Weekly Model Watch #7

September 22, 2026 · MarketMania Research · Weekly series · window Sep 14-20, 2026 UTC

RESEARCH SNAPSHOT

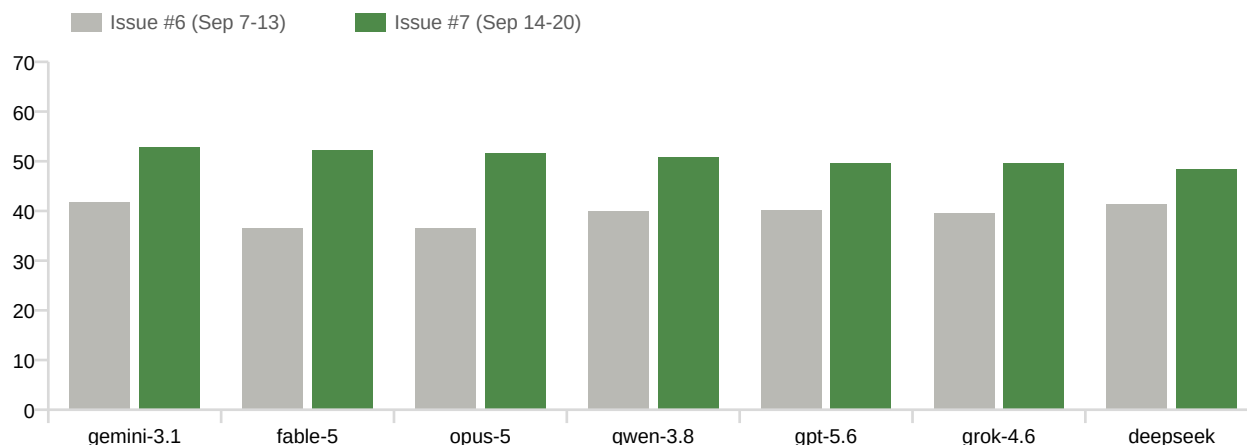
5,598 directional calls scored	7 models (stable lineup)	of 9,308 mature forecasts	window Sep 14-20, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 50.7%	cutoff Mon Sep 21, 2026, 16:00 UTC
market: BTC net +5.64% (prior -4.36%)	ann vol 51.0% (was 19.7%)	TOP5 volume \$18.74B , +17% w/w	pairwise corr 0.93 (was 0.74)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 39.6% -> 50.7% (+11.2pp) with 7 of 7 models improving, and no weekly title was awarded, for the seventh issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** gemini-3.1-pro leads descriptively at 52.9% [49.6%, 56.2%]; the gap to #2 (claude-fable-5, 52.3%) is 0.6pp. 7 of 7 models improved week over week; issue #6's leader gemini-3.1-pro is now 1 of 7.
- **1 of 5 tickers were readable.** ETH is the week's most readable coin (57.2%, n=1,116) and BNB the hardest at 46.3%; 1 of the 5 assets cleared 51%. Pair of the week: gemini-3.1-pro x ETH 57.9% (n=164); worst: qwen-3.8-max x BNB 43.6% (n=149).
- **1 of 1 qualifying turns found a caller.** ETH long +2.22%, 3 caller(s), first gemini-3.1-pro.
- **Self-agreement read as a lift in every cell.** Cross-TF/FH lifts ran +1.4pp to +6.5pp (issue #6: -6.0pp to +0.3pp). On trend days, 4h TF-agreement hit 62.9% (n=353) against 41.4% (n=304) 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 39.6% -> 50.7%. The grok column compares two pure grok-4.6 weeks: the 4.5 -> 4.6 flip sits three windows back. "Was" values are the numbers published in issue #6; 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 #6 as well -- the 4.5 -> 4.6 flip sits three 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 three 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 second 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	883	66.4%	52.9%	49.6%-56.2%	+11.1
claude-fable-5	799	60.1%	52.3%	48.9%-55.8%	+15.8
claude-opus-5	711	53.5%	51.6%	47.9%-55.3%	+15.1
qwen-3.8-max	819	61.6%	50.8%	47.4%-54.2%	+10.8
gpt-5.6-sol	821	61.7%	49.7%	46.3%-53.1%	+9.6
grok-4.6	629	47.3%	49.6%	45.7%-53.5%	+10.0
deepseek-v4-pro	936	70.4%	48.4%	45.2%-51.6%	+7.0

No weekly title is awarded this issue -- for the seventh time. Title rule: ≥ 5 pp gap AND non-overlapping 95% CI, $N \geq 10$ per side. gemini-3.1-pro's lead over claude-fable-5 is 0.6pp and their CIs overlap, so neither condition is met. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits three 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 second pure-4.6 against pure-4.6 comparison of the series.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #6
ETH	1,116	57.2%	54.2%-60.0%	29.5%
XRP	1,108	50.4%	47.4%-53.3%	42.2%
BTC	1,163	50.3%	47.4%-53.2%	41.4%
SOL	1,156	49.5%	46.6%-52.4%	36.0%
BNB	1,055	46.3%	43.3%-49.3%	46.9%

ETH was the field's most readable ticker (57.2%) and BNB the hardest (46.3%); 1 of 5 assets cleared 51% this week (issue #6: 0 of 5). Issue #6's best ticker was BNB and its hardest ETH -- this issue's ordering is ETH > XRP > BTC > SOL > 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	ETH	164	57.9%
claude-fable-5	ETH	152	57.9%
claude-opus-5	ETH	147	57.8%
grok-4.6	ETH	135	57.8%
qwen-3.8-max	ETH	160	56.9%
gpt-5.6-sol	ETH	167	56.3%

Model (5 worst)	Symbol	N	Hit rate
gpt-5.6-sol	BNB	152	45.4%
deepseek-v4-pro	SOL	192	45.3%
grok-4.6	SOL	129	45.0%
deepseek-v4-pro	BNB	171	43.9%
qwen-3.8-max	BNB	149	43.6%

gemini-3.1-pro x ETH (57.9%, n=164) is the week's best pair and qwen-3.8-max x BNB (43.6%, n=149) the weakest. Pair history accumulates across issues -- read this as the seventh 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
ETH	Sun Sep 20, 12:00	Long	+2.22%	3	gemini-3.1-pro

The ETH long turn drew 3 caller(s), first gemini-3.1-pro on a 1d/1d call at Sep 20, 00:01 with confidence 75. Scope in v1 remains BTC/ETH only; the event count across seven 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	657	53.0%	27	48.1%	51.6% (898)	+1.4
cross-TF	1d	TF 1d vs 4h	133	39.1%	3 *	100.0%	36.8% (166)	+2.4
cross-FH	4h	by FH 1h	584	53.1%	36	44.4%	51.6% (898)	+1.5
cross-FH	1d	by FH 4h	104	43.3%	7 *	57.1%	36.8% (166)	+6.5

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 214 / 30 / 278 / 55. Pooled across all four cell families the lifts run +1.4pp to +6.5pp this week (issue #6: -6.0pp to +0.3pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has 3 trend days of 7: 09-14 +1.79% (trend), 09-15 -3.35% (trend), 09-16 +0.86%, 09-17 +0.21%, 09-18 +5.82% (trend), 09-19 +0.46%, 09-20 +0.13%. On trend days, 4h TF-agreement hit 62.9% (n=353) against 41.4% (n=304) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	44.0%	16.1%	39.9%	2.73
claude-opus-5	1,330	40.0%	13.5%	46.5%	2.97
deepseek-v4-pro	1,330	43.1%	27.3%	29.6%	1.58
gemini-3.1-pro	1,330	41.3%	25.1%	33.6%	1.64
gpt-5.6-sol	1,330	34.7%	27.1%	38.3%	1.28
grok-4.6	1,329	26.3%	21.1%	52.7%	1.25
qwen-3.8-max	1,329	40.7%	20.9%	38.4%	1.95

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 1.91 this week. 'Sideways' was 39.9% of mature forecasts (3,710 of 9,308). Consensus skew: 685 symbol/FH/TF cells had 5 or more models on the same side and hit 50.5% (n=4,356) against the 50.7% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gpt-5.6-sol	BNB	Short	84	1h / 1h	Sep 16, 09:01	EXPIRY	-\$0.67

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 (52.9%); the best-calibrated line is claude-fable-5 (Brier 0.2593, gap +9.1pp). Per-model ok-rate this window ran 99.46% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #7.

Market check: an up week on a louder tape

OBSERVATION — mean |1d move| 1.35% -> 2.69% (+100% rel), BTC realized vol 31.4% -> 35.1% (ann., hourly); field directional accuracy 39.6% -> 50.7% (+11.2pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -\$1018.93 -> -\$337.72 (adjacent calendar weeks Sep 7-13 vs Sep 14-20; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 38.7% -> 54.2% — the same direction as the trade-based hit rule.

Measure	Sep 7-13 (week A)	Sep 14-20 (week B)	Change
Mean 1d move (5 assets)	1.35%	2.69%	+100% rel
BTC realized vol (ann., hourly)	31.4%	35.1%	+3.7pp
Field directional accuracy	39.6%	50.7%	+11.2pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	38.7%	54.2%	+15.4pp
Field sim win-rate	32.1%	43.4%	+11.2pp
Field sim net PnL	-\$1018.93	-\$337.72	--

Week A is exactly the issue-#6 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, three windows back, inside issue #4's window). Week A (2026-09-07..2026-09-13) is pure grok-4.6 (it reproduces the published issue #6); 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. Seven issues, and the best ticker changed hands in every one except issue #6 (SOL -> BNB -> BTC -> SOL -> BNB -> BNB -> ETH): ETH leads this week at 57.2%, after two issues of BNB.
- Self-agreement lifts run +1.4pp to +6.5pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Sep 14-20, 2026 UTC). Descriptive for this window only; no claim about next week or any model's underlying skill. This window ran 3 trend days of 7 (issue #6: 2 of 7); every comparison with issue #6 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. 13 events across seven issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=27, 3, 36, 7); 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 the retirement of V4 Pro in favour of V4.1 Flash from Sep 14 (Beijing noon); as of Sep 22 the API still returns deepseek-v4-pro as the served model for our requests and the per-answer latency, token and cost series of Sep 1-22 show no regime change, so this issue keeps the label; it will change the day the served id changes. Series density and lineage notes (wave 7): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 14-20) daily

coverage is FULL, as it was in issues #4, #5 and #6: 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, three 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-#6 window), so every grok number in this issue and in issue #6 is a pure grok-4.6 line -- the grok week-over-week row is the second 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 #6 after the audit that found the raw table mixes two exchanges; issue #6 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 21 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5 and #6, which printed 47,373 at the Sep 14 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 52,971 is an additive step under one definition; counter deltas against issue #5 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,308
OK in gate	9,308	-- of them directional	5,598
Out of gate (1w / 1M)	488 / 485	-- of them sideways	3,710
Invalid	9	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-14.json	Report cutoff	Mon Sep 21, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-22 06:56 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,598 scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 21, 2026 cutoff): 52,971 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 33 published reports.

Issue #7. Weekly Model Watch is a living series. Issue #6's open questions -- does gemini-3.1-pro hold the top rank for a third issue and does the title rule ever fire, and does a qualifying reversal ever fire outside ETH or a second turn in the same pair draw a caller? -- read this issue as: gemini-3.1-pro at the top (52.9%), the line that also led issue #6 (gemini-3.1-pro), so the rank held for a third issue; no weekly title was awarded, for the seventh issue running, and 1 of 1 qualifying turns found a caller, on turns in ETH only. The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #6; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits three 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 fourth issue, and does the title rule ever fire?
- Does a qualifying reversal ever fire outside ETH, 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 #7	https://marketmania.ai/research/reports/consensus-watch-2026-09-14.pdf
Weekly Calibration #7	https://marketmania.ai/research/reports/weekly-calibration-2026-09-14.pdf
Weekly Model Watch #6	https://marketmania.ai/research/reports/model-watch-2026-09-07.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_2026w38,
  title = {Weekly Model Watch #7: weekly leaderboard, ticker/pair reads and cross-confirmation, Sep
14-20 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {September}, day = {22},
  url   = {https://marketmania.ai/research/reports/model-watch-2026-09-14.pdf},
  note  = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Sep 14-20, 2026 UTC; source
weekly_metrics_2026-09-14.json}
}
```