

Consensus Watch #5

September 8, 2026 · MarketMania Research · Weekly series · window Aug 31-Sep 6, 2026 UTC

RESEARCH SNAPSHOT

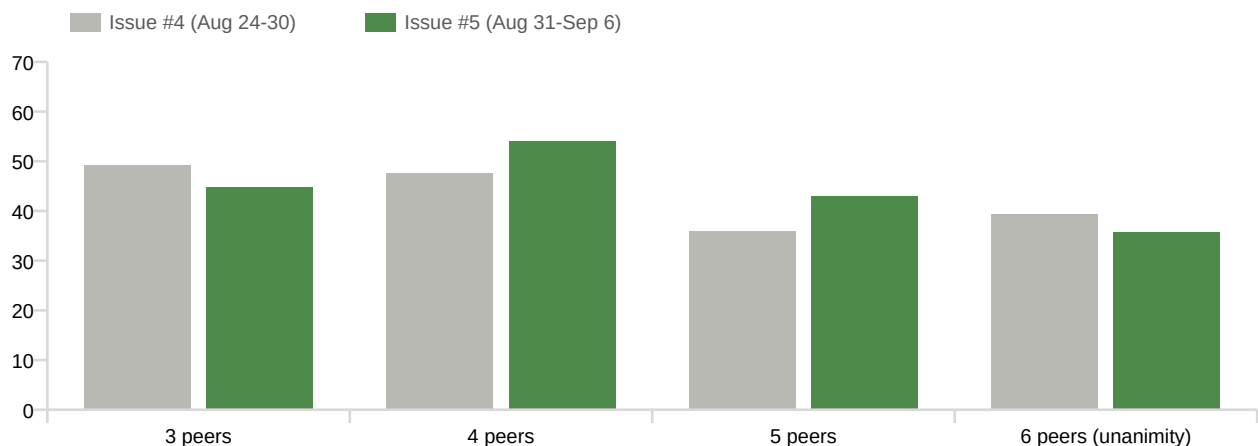
3,751 LOO-scoreable calls	7 models (stable lineup)	769 thin/tied excluded	window Aug 31-Sep 6, 2026 UTC
method = leave-one-out majority, 3+ directional peers	FH gate 1h / 4h / 1d	base field hit 42.8%	cutoff Mon Sep 7, 2026, 16:00 UTC
market: BTC net +3.42% (prior -0.07%)	ann vol 43.4% (was 30.9%)	TOP5 volume \$16.24B , -18% w/w	pairwise corr 0.78 (was 0.85)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 99.3% of leave-one-out scoreable calls and the pooled lift came in at -27.6pp: agree-hit 41.6% against 69.2% on disagreement, on a disagree bucket of n=26.

TL;DR

- **OBSERVATION -- 99.3% of the field's leave-one-out calls ran with the herd.** Of 3,751 LOO-scoreable directional calls, 3,725 (99.3%) sided with peers' majority (issue #4: 99.1%); agree-hit was 41.6% [40.0%, 43.2%] vs. 69.2% [50.0%, 83.5%] for the 26 disagreeers, a pooled lift of -27.6pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a fifth reading.** 3 peers 44.7% (was 49.3%), 4 peers 54.0% (was 47.6%), 5 peers 43.0% (was 35.9%), full unanimity 35.7% (was 39.4%). 3 of the 4 tiers sit above the 42.8% field base, and 3 of the 4 kept the side of the base they took in issue #4.
- **Daily-horizon unanimity fell below the field base.** 1d/1d margin-6 hit 12.5% (n=56) and 1d/4h margin-6 hit 21.4% (n=42); issue #4 printed 76.2% and 50.5% on the same two cells.
- **Herd failures did not disappear.** 392 cells had 6+ directional models on one side (issue #4: 408); the 5 worst sat at 67.9-68.6 mean confidence and resolved 0.0%-28.6%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #5 n: 456 / 615 / 918 / 1,715; field base this week 42.8% (issue #4: 42.9%). A 2-peer bucket exists this week too (n=21, 23.8%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #4.

Why it matters

MarketMania publishes forecasts from 7 frontier model lines into the same slots every week; a natural question for anyone reading the feed is whether a model's call is worth more when it agrees with what everyone else is saying, or whether that agreement is just everyone making the same mistake together. Consensus Watch answers that question empirically, one week at a time, using a leave-one-out design built specifically to avoid the obvious trap of comparing a model to an index that includes its own vote.

How to read this

For every model M and every forecast it made, we rebuild that slot's consensus using only the OTHER active models in the same symbol / forecast-horizon (FH) / timeframe (TF) cell -- M's own call never counts toward its own consensus. M is scored agree if its side (long or short) matches the strict majority of those peers, and disagree if it sits alone against that majority; a cell needs at least 3 directional peers to count at all, and tied peer splits are excluded rather than forced either way. LOO peers are genuinely external to the model being scored, so the comparison below is not comparing a model to itself. The grok rows are one line, so grok never counts as two peers in a cell. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits inside the PREVIOUS window (week A, issue #4): every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)"; this window is pure grok-4.6, so the grok week-over-week row compares a pure-4.6 week with a spliced week.

Pooled result: agree vs. disagree with the LOO consensus

Group	n	Hit rate	Wilson 95% CI	Lift, pp
Agree with LOO peers	3,725	41.6%	[40.0%, 43.2%]	-27.6 (vs. disagree)
Disagree with LOO peers	26	69.2%	[50.0%, 83.5%]	ref.

Read with context. The disagree bucket is n=26 this week against 36 in issue #4 and 125 in issue #3; the two Wilson intervals do not overlap. This is a single week of dependent observations: not evidence about herding in general, and not a claim that the sign will hold next week.

By forecast horizon / timeframe

The pooled result above mixes 5 different FH/TF cells. Broken out:

FH / TF	Agree n	Agree hit	Dis. n	Dis. hit	Lift, pp	Excluded
1h / 1h	2,402	44.7%	20	65.0%	-20.3	503
4h / 4h	553	38.5%	3	100.0%	-61.5	95
4h / 1h	558	33.5%	3	66.7%	-33.2	131
1d / 1d	98	37.8%	0	n/a	n/a	30
1d / 4h	114	34.2%	0	n/a	n/a	10

"Excluded" = thin (fewer than 3 directional peers) or tied peer splits in that cell; not folded into agree or disagree. "Dis." = disagree. Cells whose disagree bucket is under N=10 are insufficient and are never used to rank anything.

Margin: does more agreement mean a safer call?

For every LOO-scored forecast we also count how many peers were on the winning (majority) side -- from the bare minimum of 3 up to the full peer set agreeing (6 of 6, since 7 model lines were active and LOO always drops one). The week-over-week chart is on page 1; the full per-tier numbers: 3 peers 44.7% (n=456), 4 peers 54.0% (n=615), 5 peers 43.0% (n=918), 6 peers 35.7% (n=1,715). Against the 42.8% field base, 3 tier(s) sit above and 1 below, and 3 of the 4 tiers kept the side of the base they took in issue #4. Five issues in, the margin curve has taken five readings.

Daily-horizon herds: unanimity below the field base

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 12.5% (n=56) and 1d/4h unanimity 21.4% (n=42), both below the 42.8% field base; the whole 1d/1d agree bucket ran 37.8% (n=98) -- not the highest of the five FH/TF cells -- while 1d/4h ran 34.2% (n=114), the second-thinnest agree bucket in the table above. Issue #4 printed 76.2% and 50.5% on the same two unanimity cells. Both daily buckets are an order of magnitude thinner than the hourly one (n=2,402), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 392 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #4: 408).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Sep 2, 02:01	BTC	1h / 1h	short	7	68.6	0.0%
Sep 6, 12:01	SOL	4h / 1h	long	7	68.4	0.0%
Sep 6, 13:01	SOL	1h / 1h	long	7	68.3	28.6%
Sep 3, 16:01	SOL	4h / 4h	long	7	68.1	14.3%
Sep 2, 02:01	XRP	1h / 1h	short	7	67.9	0.0%

The 5 worst cells were long/short swarms at 67.9-68.6 mean confidence and resolved 0.0%-28.6%. 'Sideways' was 51.2% of mature forecasts (4,744 of 9,264) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 0 to 11 this week (issue #4: 0 to 23). 1 of the 7 lines clear the N=10 floor (gemini-3.1-pro 11); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	557	42.5%	1	0.0%
claude-opus-5	489	42.3%	2	50.0%
deepseek-v4-pro	560	42.7%	6	83.3%
gemini-3.1-pro	554	44.2%	11	72.7%
gpt-5.6-sol	572	41.4%	0	n/a
grok-4.6	428	36.0%	4	50.0%
qwen-3.8-max	565	40.7%	2	100.0%

Individual per-model contrarian records are anecdotes, not findings: at these counts a handful of flipped calls swings a 'lift' by tens of points, so models are never ranked by contrarian lift here. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits inside the PREVIOUS window (week A, issue #4): every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)"; this window is pure grok-4.6, so the grok week-over-week row compares a pure-4.6 week with a spliced week.

Market check: an up week on lower volume

OBSERVATION — mean |1d move| 2.35% -> 2.10% (-11% rel), BTC realized vol 37.9% -> 35.2% (ann., hourly); field directional accuracy 42.9% -> 42.8% (-0.1pp), 1 of 7 models improved, sim win-rate up for 1 of 7, field sim PnL -\$473.59 -> -\$669.03 (adjacent calendar weeks Aug 24-30 vs Aug 31-Sep 6; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 44.2% -> 51.6% — the opposite direction to the trade-based hit rule.

Measure	Aug 24-30 (week A)	Aug 31-Sep 6 (week B)	Change
Mean 1d move (5 assets)	2.35%	2.10%	-11% rel
BTC realized vol (ann., hourly)	37.9%	35.2%	-2.7pp
Field directional accuracy	42.9%	42.8%	-0.1pp
Models improving hit-rate	--	1 of 7	--
Raw price-sign accuracy	44.2%	51.6%	+7.4pp
Field sim win-rate	37.0%	35.6%	-1.4pp
Field sim net PnL	-\$473.59	-\$669.03	--

Week A is exactly the issue-#4 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, inside week A). Week A (2026-08-24..2026-08-30) mixes the 4.5-era (Aug 24 00:00-09:22) with 4.6 and reproduces the published issue #4; week B is pure grok-4.6.

Practical implications

- Five issues, five margin-curve readings. The tier ordering moved again: 4 peers went 47.6% -> 54.0% while full unanimity went 39.4% -> 35.7%, so a consensus filter still should not carry weight in a config.
- The pooled lift is -27.6pp this week against -6.0pp in issue #4 and +17.1pp in issue #3. Three readings, three magnitudes: treat the number as this week's weather, not an edge.
- Read the consensus number together with the market check on the same page. The field base moved -0.1pp week over week; 2 of the 4 consensus tiers moved up and 2 down against issue #4 (3 peers -4.6pp, 4 peers +6.4pp, 5 peers +7.1pp, unanimity -3.7pp).

Limitations

- Single week (Aug 31-Sep 6, 2026 UTC). Every finding is descriptive for this window only.
- Observations are not independent: the same models watch overlapping symbol / FH / TF cells hour after hour. Wilson intervals here are descriptive, not inferential.
- Correlational, not causal: all models see the same market data, so 'agreement' and 'hit' can rise together simply because a slot was easy to read. LOO removes self-match bias, not this confound.
- The pooled lift rests on 26 disagreeing calls; per-model counts run 0-11. 769 calls (17.0% of the 4,520-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window ran 3 trend days of 7 (issue #4: 3 of 7); every comparison with issue #4 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 5): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 31-Sep 6) daily coverage is FULL, as it was in issue #4 (the first full week of the series): 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 490). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, inside the PREVIOUS window (week A, issue #4): this window carries 1,470 grok rows and none from the 4.5 era, so every grok number in this issue is a pure grok-4.6 line, while every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)" -- the grok week-over-week row compares a pure-4.6 week with a spliced week. (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 #4 after the audit that found the raw table mixes two exchanges; issue #4 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 7 16:00 UTC) -- the definition pinned in issue #3 and carried by issue #4, which printed 38,385 at the Aug 31 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 42,905 is an additive step under one definition; counter deltas against issue #3 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,264
OK in gate	9,264	-- of them directional	4,520
Out of gate (1w / 1M)	490 / 490	-- of them sideways	4,744
Invalid	46	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-08-31.json	Report cutoff	Mon Sep 7, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-08 08:08 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 3,751 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 7, 2026 cutoff): 42,905 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 23 published reports.

Issue #5. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #4 asked whether the pooled agree-vs-disagree sign settles: it reads -27.6pp this week after -6.0pp in issue #4 (agree 41.6% vs disagree 69.2% on n=26), and whether any margin tier repeats its position against the field base two issues running: 3 of the 4 tiers did, and the 4-peer tier came in above the field base. The grok row is pure grok-4.6 this issue; issue #4's grok row was the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)". 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 the pooled agree-vs-disagree sign keep the same direction for a third issue running?
- Do the same margin tiers hold their side of the field base for a third issue?
- Monthly series: the agreement curve across regimes at monthly n.

Related research

Report	Direct PDF link
Weekly Calibration #5	https://marketmania.ai/research/reports/weekly-calibration-2026-08-31.pdf
Weekly Model Watch #5	https://marketmania.ai/research/reports/model-watch-2026-08-31.pdf
Consensus Watch #4	https://marketmania.ai/research/reports/consensus-watch-2026-08-24.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_consensus_2026w36,
  title = {Consensus Watch #5: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {8},
  url = {https://marketmania.ai/research/reports/consensus-watch-2026-08-31.pdf},
  note = {Methodology v1.1; window Aug 31-Sep 6, 2026 UTC; source weekly_metrics_2026-08-31.json}
}
```