

Consensus Watch #4

September 2, 2026 · MarketMania Research · Weekly series · window Aug 24-30, 2026 UTC

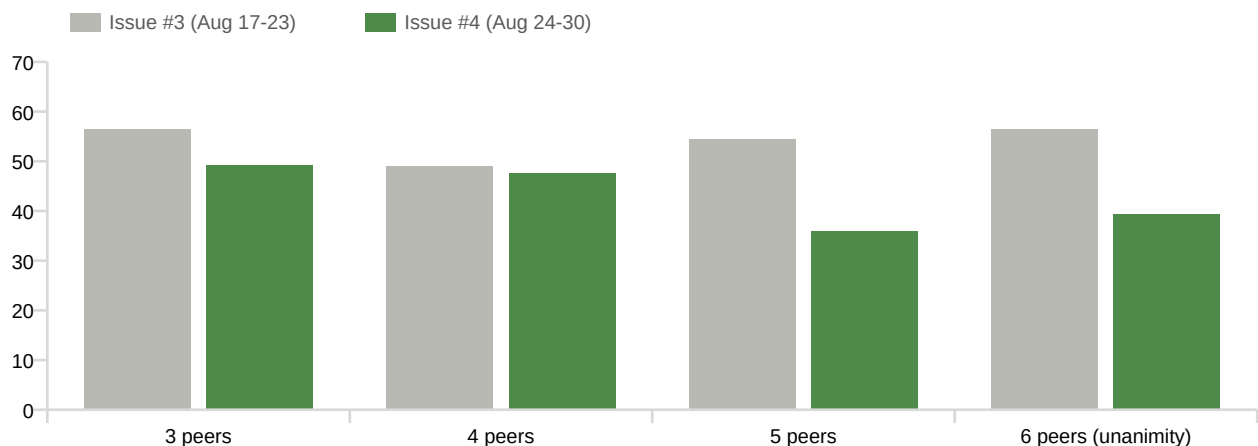
RESEARCH SNAPSHOT			
3,884 LOO-scoreable calls	7 models (grok line spliced)	692 thin/tied excluded	window Aug 24-30, 2026 UTC
method = leave-one-out majority, 3+ directional peers	FH gate 1h / 4h / 1d	base field hit 42.9%	cutoff Mon Aug 31, 2026, 16:00 UTC
market: BTC net -0.07% (prior +23.58%)	ann vol 30.9% (was 71.1%)	TOP5 volume \$19.72B , -24% w/w	pairwise corr 0.85 (was 0.81)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 99.1% of leave-one-out scoreable calls and the pooled lift came in at -6.0pp: agree-hit 41.2% against 47.2% on disagreement, on a disagree bucket of n=36.

TL;DR

- **OBSERVATION -- 99.1% of the field's leave-one-out calls ran with the herd.** Of 3,884 LOO-scoreable directional calls, 3,848 (99.1%) sided with peers' majority (issue #3: 97.1%); agree-hit was 41.2% [39.6%, 42.8%] vs. 47.2% [32.0%, 63.0%] for the 36 disagreeers, a pooled lift of -6.0pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a fourth shape.** 3 peers 49.3% (was 56.4%), 4 peers 47.6% (was 48.9%), 5 peers 35.9% (was 54.4%), full unanimity 39.4% (was 56.5%). 2 of the 4 tiers sit above the 42.9% field base; four issues in, the curve has still not repeated a shape.
- **Daily-horizon unanimity stayed above the field base.** 1d/1d margin-6 hit 76.2% (n=21) and 1d/4h margin-6 hit 50.5% (n=105); issue #3 printed 81.0% and 61.9% on the same two cells.
- **Herd failures did not disappear.** 408 cells had 6+ directional models on one side (issue #3: 437); the 5 worst sat at 68.0-68.9 mean confidence and resolved 0.0%-42.9%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #4 n: 456 / 620 / 924 / 1,827; field base this week 42.9% (issue #3: 54.1%). A 2-peer bucket exists this week too (n=21, 61.9%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #3.

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 lineage-spliced line this issue (grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)), so grok never counts as two peers in a cell.

Pooled result: agree vs. disagree with the LOO consensus

Group	n	Hit rate	Wilson 95% CI	Lift, pp
Agree with LOO peers	3,848	41.2%	[39.6%, 42.8%]	-6.0 (vs. disagree)
Disagree with LOO peers	36	47.2%	[32.0%, 63.0%]	ref.

Read with context. The disagree bucket is n=36 this week against 125 in issue #3 and 6 in issue #2; the two Wilson intervals 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	2,374	40.4%	26	42.3%	-1.9	462
4h / 4h	599	41.6%	4	75.0%	-33.4	109
4h / 1h	601	41.4%	5	40.0%	+1.4	85
1d / 1d	115	53.0%	1	100.0%	-47.0	27
1d / 4h	159	41.5%	0	n/a	n/a	9

"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 49.3% (n=456), 4 peers 47.6% (n=620), 5 peers 35.9% (n=924), 6 peers 39.4% (n=1,827). Against the 42.9% field base, 2 tier(s) sit above and 2 below. Four issues in, the margin curve has taken four readings and no tier has yet carried a stable meaning.

Daily-horizon herds: the thinnest cell, the best rate

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 76.2% (n=21) and 1d/4h unanimity 50.5% (n=105), both above the 42.9% field base; the whole 1d/1d agree bucket ran 53.0% (n=115) -- the highest of the five FH/TF cells -- while 1d/4h ran 41.5% (n=159), in line with the hourly and 4h cells. Issue #3 printed 81.0% and 61.9% on the same two unanimity cells. Both daily buckets are an order of magnitude thinner than the hourly one (n=2,374), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 408 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #3: 437).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Aug 26, 12:01	XRP	1h / 1h	short	7	68.9	42.9%
Aug 27, 15:01	SOL	1h / 1h	long	7	68.9	0.0%
Aug 27, 15:01	XRP	1h / 1h	long	7	68.7	0.0%
Aug 24, 16:01	SOL	4h / 4h	long	6	68.3	0.0%
Aug 25, 04:01	SOL	4h / 1h	long	7	68.0	0.0%

The 5 worst cells were long/short swarms at 68.0-68.9 mean confidence and resolved 0.0%-42.9%. 'Sideways' was 50.8% of mature forecasts (4,725 of 9,301) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 0 to 23 this week (issue #3: 2 to 53). 1 of the 7 lines clear the N=10 floor (gemini-3.1-pro 23); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	572	41.4%	1	100.0%
claude-opus-5	531	42.2%	0	n/a
deepseek-v4-pro	572	42.5%	6	16.7%
gemini-3.1-pro	544	40.1%	23	52.2%
gpt-5.6-sol	610	40.8%	1	100.0%
grok-4.6 (spliced)	451	40.1%	4	50.0%
qwen-3.8-max	568	41.0%	1	0.0%

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. grok-4.6 (spliced) = grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22).

Market check: the tape cooled off

OBSERVATION — mean |1d move| 4.28% -> 2.35% (-45% rel), BTC realized vol 58.7% -> 37.9% (ann., hourly); field directional accuracy 54.0% -> 42.9% (-11.1pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL +\$889.37 -> -\$473.59 (adjacent calendar weeks Aug 17-23 vs Aug 24-30; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 53.5% -> 44.2% — the same direction as the trade-based hit rule.

Measure	Aug 17-23 (week A)	Aug 24-30 (week B)	Change
Mean 1d move (5 assets)	4.28%	2.35%	-45% rel
BTC realized vol (ann., hourly)	58.7%	37.9%	-20.8pp
Field directional accuracy	54.0%	42.9%	-11.1pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	53.5%	44.2%	-9.3pp
Field sim win-rate	46.3%	37.0%	-9.2pp
Field sim net PnL	+\$889.37	-\$473.59	--

Week A is exactly the issue-#3 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 B). Week A (2026-08-17..2026-08-23) is pure grok-4.5 and reproduces the published issue #3; week B mixes the 4.5-era (Aug 24 00:00-09:22) with 4.6.

Practical implications

- Four issues, four margin-curve readings. No consensus tier has shown a stable meaning, so a consensus filter still should not carry weight in a config.
- The pooled lift is -6.0pp this week against +17.1pp in issue #3 and -6.9pp in issue #2. 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 -11.1pp week over week, and every consensus tier moves with it.

Limitations

- Single week (Aug 24-30, 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 36 disagreeing calls; per-model counts run 0-23. 692 calls (15.1% of the 4,576-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window ran 3 trend days of 7 (issue #3: 5 of 7); every comparison with issue #3 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 4): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 24-30) daily coverage is FULL for the first time in 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 488, 1M 490). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, INSIDE this window: every grok row in this issue is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)" (110 rows from the 4.5 era, 1,360 from 4.6); issue #3 was pure grok-4.5, so the grok week-over-week row joins a spliced line to a single-model line. (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 #3 after the audit that found the raw table mixes two exchanges; issue #3 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 (Aug 31 16:00 UTC) -- the definition pinned in issue #3, which printed 33,809 at the Aug 24 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 38,385 is an additive step under one definition; counter deltas against issue #2 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,301
OK in gate	9,301	-- of them directional	4,576
Out of gate (1w / 1M)	488 / 490	-- of them sideways	4,725
Invalid	11	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-08-24.json	Report cutoff	Mon Aug 31, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-02 06:14 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 3,884 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Aug 31, 2026 cutoff): 38,385 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 15 published reports.

Issue #4. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #3 asked whether the agree-side lift survives a fourth week: this week it reads -6.0pp (agree 41.2% vs disagree 47.2% on n=36), and whether the 4-peer dip repeats: the 4-peer tier came in above the field base. The grok line is spliced this issue (grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)) because the 4.5 -> 4.6 flip happened inside the window. 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 settle, or does it keep changing with the regime?
- Does any margin tier repeat its position against the field base two issues running?
- Monthly series: the agreement curve across regimes at monthly n.

Related research

Report	Direct PDF link
Weekly Calibration #4	https://marketmania.ai/research/reports/weekly-calibration-2026-08-24.pdf
Weekly Model Watch #4	https://marketmania.ai/research/reports/model-watch-2026-08-24.pdf
Consensus Watch #3	https://marketmania.ai/research/reports/consensus-watch-2026-08-17.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_2026w35,
  title = {Consensus Watch #4: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {2},
  url = {https://marketmania.ai/research/reports/consensus-watch-2026-08-24.pdf},
  note = {Methodology v1.1; window Aug 24-30, 2026 UTC; source weekly_metrics_2026-08-24.json}
}
```