

Consensus Watch #2

August 19, 2026 · window Aug 10-16 · MarketMania Research · Weekly series

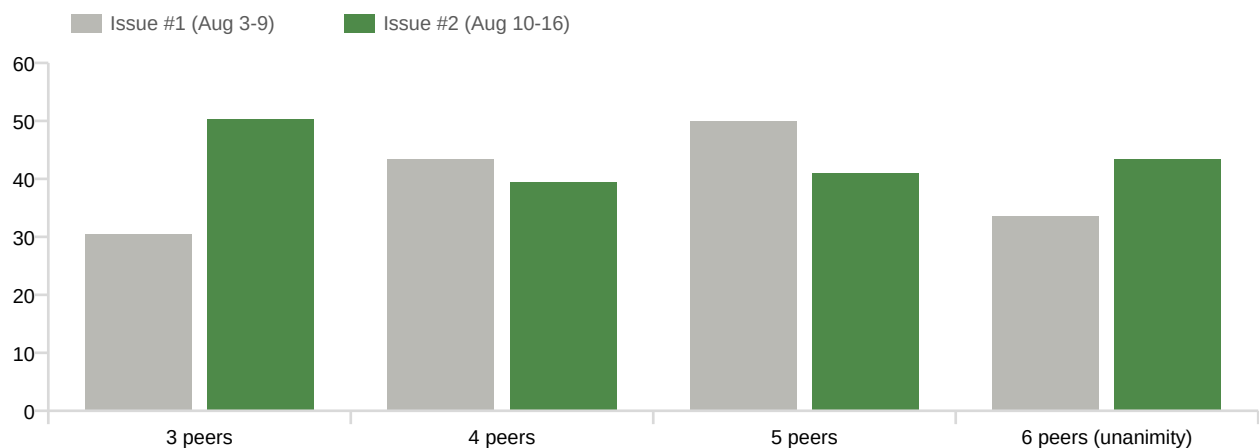
RESEARCH SNAPSHOT			
2,929 LOO-scoreable calls	7 models (stable lineup)	741 thin/tied excluded	window Aug 10-16, 2026 UTC
method = leave-one-out majority	3+ directional peers required	base field hit 43.8%	cutoff Mon Aug 17, 2026, 16:00 UTC
market: BTC net -3.08% (prior +2.09%)	ann. vol 10.0% (was 11.4%)	TOP5 volume \$7.75B , -8.9% w/w	pairwise corr 0.52 (was 0.40)

KEY FINDING · OBSERVATION (one weekly window)

The margin curve inverted this week: the thinnest majorities (3 peers) were the only tier to beat the field, and near-unanimity stayed below base for a second week.

TL;DR

- **OBSERVATION -- Herding intensified, and still did not pay.** Of 2,929 LOO-scoreable directional calls, 2,923 (99.8%) sided with peers' majority (issue #1: 99.4%); agree-hit was 43.1% vs. 50.0% for the 6 disagreeers, a pooled lift of -6.9pp on n=6 with a Wilson 95% CI of [18.8%, 81.2%] that swallows the agree bucket's own CI [41.3%, 44.9%]. Descriptive, one week, not a claim of a durable edge.
- **The margin curve flipped upside-down vs. issue #1.** Hit rate by agreeing peers ran 3: **50.2%** (was 30.5%), 4: 39.4% (was 43.3%), 5: 41.0% (was 49.9%), full unanimity: 43.3% (was 33.6%). Only the thinnest majority beat the 43.8% base. Two weeks, two opposite shapes -- the tier ordering itself looks unstable, which is a finding about consensus signals, not a trading edge.
- **Daily-horizon herds were the week's disaster.** On 1d/1d cells, margin-5 and margin-6 consensus hit 19.4% and 17.1%; on 1d/4h, near-unanimous calls went **0 for 28**. Intraday herds were mediocre; daily herds were outright wrong.
- **Herd failures were real.** 299 cells had 6+ directional models on the same side; the week's 5 worst were total wipeouts -- all 7 models long at ~70 mean confidence, 0% hit -- in a week where 'sideways' was the field's most common call (60.6% of mature forecasts).



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #2 n: 428 / 525 / 792 / 1,169; field base this week 43.8%. The tier that led last week (5 peers) fell 8.9pp; the tier that trailed (3 peers) now leads at 50.2%. A 2-peer bucket exists this week too (n=9, 66.7%) -- footnote only, far too small to plot.

Why it matters

MarketMania publishes forecasts from 7 frontier models 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. This leave-one-out (LOO) design exists because the obvious alternative -- comparing a model to a consensus that includes its own vote -- is biased in the model's favor: with the model counted inside the index, its own vote nudges that index toward itself, mechanically inflating the agree bucket's measured hit rate and the resulting lift. LOO peers are genuinely external to the model being scored, so the comparison below is not comparing a model to itself.

Pooled result: agree vs. disagree with the LOO consensus

Group	n	Hit rate	Wilson 95% CI	Lift, pp
Agree with LOO peers	2,923	43.1%	[41.3%, 44.9%]	-6.9 (vs. disagree)
Disagree with LOO peers	6	50.0%	[18.8%, 81.2%]	ref.

Read with caution. The -6.9pp pooled lift above rests on just 6 disagreeing calls -- a third of issue #1's already-thin 19. Its Wilson 95% CI, [18.8%, 81.2%], swallows the agree bucket's CI entirely. This is a descriptive, single-week result -- not evidence that contrarian calls are more reliable, and not a claim that will necessarily hold up next week.

By forecast horizon / timeframe

The pooled result above mixes 5 different FH/TF cells. Broken out, LOO disagreement again only happened in the FH 1h / TF 1h cell (all 6 calls); the other four cells had zero LOO-disagreeing directional calls -- for the second week running, every model with a valid majority outside the hourly cell simply went along with it.

FH / TF	Agree n	Agree hit	Dis. n	Dis. hit	Lift, pp	Excluded
1h / 1h	1,865	45.0%	6	50.0%	-7.0	475
4h / 4h	435	40.9%	0	n/a	n/a	142
4h / 1h	463	40.8%	0	n/a	n/a	90
1d / 1d	88	34.1%	0	n/a	n/a	11
1d / 4h	72	30.6%	0	n/a	n/a	23

"Excluded" = thin (fewer than 3 directional peers) or tied peer splits in that cell; not folded into agree or disagree. "Dis." = disagree.

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 models were active and LOO always drops one). The week-over-week chart appears on page 1; the full per-tier numbers: 3 peers 50.2% (n=428), 4 peers 39.4% (n=525), 5 peers 41.0% (n=792), full unanimity 43.3% (n=1,169). Only the 3-peer tier beat the week's 43.8% base rate; last week only the 4- and 5-peer tiers did. Issue #1 asked whether unanimity would keep underperforming: it did (43.3% vs. 43.8% base), but the sharper result is that the whole tier ordering flipped -- last week's best tier (5 peers, 49.9%) fell to 41.0%, and last week's worst (3 peers, 30.5%) now leads at 50.2%. Two weeks in, the margin curve carries no stable shape.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 299 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side -- effectively the whole active peer set agreeing before the outcome was known. The 5 worst of those herds were total wipeouts: every one of the 7 active models went the same direction, and the hit rate was zero.

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Aug 11, 12:01	BNB	4h / 1h	long	7	71.0	0.0%
Aug 12, 12:01	SOL	1h / 1h	long	7	69.0	0.0%
Aug 11, 13:01	BNB	1h / 1h	long	7	68.1	0.0%
Aug 11, 22:01	SOL	1h / 1h	long	7	68.0	0.0%
Aug 12, 12:01	ETH	4h / 4h	long	7	67.9	0.0%

All 5 worst herds were long calls, in a week where the field's own forecast mix leaned toward 'sideways' 60.6% of the time (5,637 of 9,307 mature forecasts) -- a 7-model long swarm was already a directional outlier before it missed. All five landed inside the same two sessions, Aug 11-12.

Per-model: agree vs. disagree

No model was immune to the herd. Every model's count of LOO-disagreeing calls this week is 0 to 4 out of several hundred agree calls each -- which is exactly why models are not ranked by individual agree/disagree lift here; at these n, a single flipped call swings a 'lift' by tens of points.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	401	41.9%	0	n/a
claude-opus-5	334	41.3%	0	n/a
deepseek-v4-pro	371	42.0%	2	0.0%
gemini-3.1-pro	406	44.6%	4	75.0%
gpt-5.6-sol	477	40.9%	0	n/a
grok-4.5	482	45.4%	0	n/a
qwen-3.8-max	452	44.7%	0	n/a

gemini-3.1-pro's 4 contrarian calls hit 3 of 4 and deepseek-v4-pro's 2 both missed; both are anecdotes, not findings. Five of seven models did not disagree with their peers even once all week.

Practical implications

- Last week unanimity behaved like a crowding signal; this week it merely matched base while the whole tier ordering flipped. The practical read is not 'fade the herd' -- it is that NO consensus tier has yet shown a stable meaning, so a consensus filter should not carry weight in a config on two weeks of evidence.
- Treat any agree-vs-disagree lift as regime-dependent description, not strategy: this week's -6.9pp rests on 6 disagreeing calls whose CI swallows the agree bucket's entirely.
- Consensus Watch is a living series: each issue adds one more week to the LOO comparison and one more point on the margin curve. Two points cannot separate shape from noise; several issues can.

Limitations

- Single week (Aug 10-16, 2026 UTC). Every finding is descriptive for this window only; no claim is made about next week or any model's underlying skill.
- Observations are not independent: the same models watch overlapping symbol / FH / TF cells hour after hour, so LOO calls made close in time are correlated, not i.i.d. draws. Wilson intervals here are descriptive, not inferential.
- The design is 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 6 disagreeing calls, all of them in FH 1h; per-model disagree counts run 0-4, far too small to rank models by contrarian lift.
- 741 calls (20.2% of the 3,670-call mature directional universe) were thin (fewer than 3 directional peers) or tied and are excluded from every consensus table -- not folded into agree or disagree either way.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	9,380	Mature (scored pool)	9,307
OK in gate	9,307	-- of them directional	3,670
Out of gate (1w)	69	Pending (next issue)	0
Invalid	4	Late closes	0
Source file	weekly_metrics_2026-08-10.json	Report cutoff	Mon Aug 17, 16:00 UTC
Generated at (pipeline)	2026-08-18 08:37 UTC	Methodology	v1.1 · hash e66c7e8c864a2233

THIS REPORT: 2,929 LOO-scored observations. **MARKETMANIA RESEARCH TO DATE** (as of Aug 17, 2026 cutoff): 23,458 resolved forecasts since Jul 11 · 9 models tracked (7 current frontier + 2 archived legacy) · 5 assets · 5 forecast horizons · hourly cadence · 9 published research reports incl. this wave.

Issue #2. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #1's open question -- does full unanimity keep underperforming? -- closed as 'yes, but barely' (43.3% vs. the 43.8% base); the sharper result is the inverted margin curve. Engine 1.1 powers the sandbox since Aug 18 (after this window closed); no cross-engine PnL comparisons are claimed.

What we're testing next

- **Next issue:** does the margin-curve inversion persist -- or is tier ordering week-to-week noise? Three points make a trend.
- **Monthly test (Sep 2):** the agreement curve across regimes and across the full month, where the daily-horizon buckets finally reach a readable n.

Related research

Weekly Calibration #2 -- <https://marketmania.ai/research/reports/weekly-calibration-2026-08-10.pdf>

Weekly Model Watch #2 -- <https://marketmania.ai/research/reports/model-watch-2026-08-10.pdf>

Config Watch #2 -- <https://marketmania.ai/research/reports/config-watch-2026-08-12.pdf>

Consensus Watch #1 -- <https://marketmania.ai/research/reports/consensus-watch-2026-08-03.pdf>

These three weekly reports publish together as one issue each week; Config Watch follows on its own cycle.

CITE THIS REPORT

```
@misc{marketmania2026cw2consensus,
  title = {Consensus Watch #2: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year  = {2026}, month = {August}, day = {19},
  url   = {https://marketmania.ai/research/reports/consensus-watch-2026-08-10.pdf},
  note  = {Methodology v1.1; window Aug 10-16, 2026 UTC; source weekly_metrics_2026-08-10.json}
}
```