

Consensus Watch #3

August 26, 2026 · MarketMania Research · Weekly series · window Aug 17-23, 2026 UTC

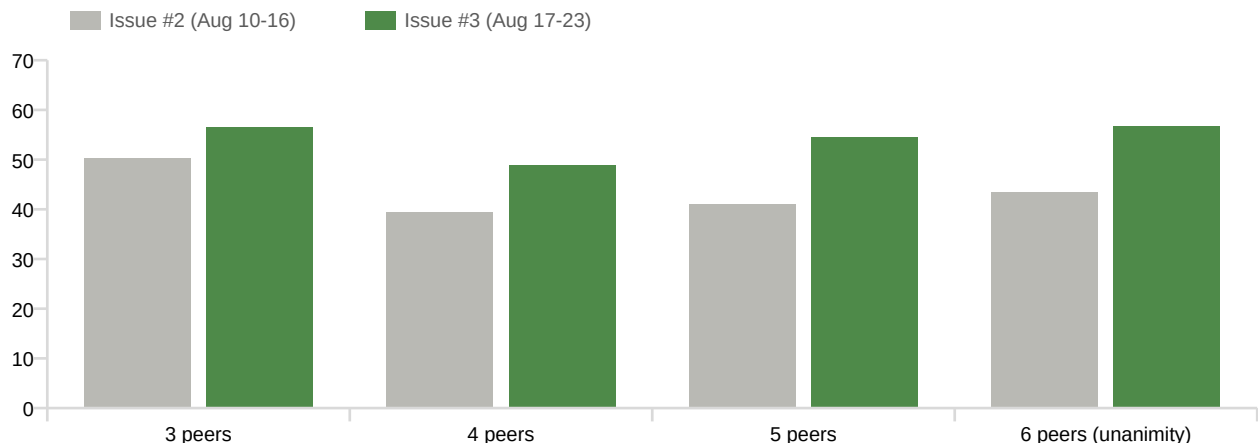
RESEARCH SNAPSHOT			
4,357 LOO-scoreable calls	7 models (stable lineup)	773 thin/tied excluded	window Aug 17-23, 2026 UTC
method = leave-one-out majority, 3+ directional peers	FH gate 1h / 4h / 1d	base field hit 54.1%	cutoff Mon Aug 24, 2026, 16:00 UTC
market: BTC net +23.58% (prior -3.08%)	ann vol 71.1% (was 10.0%)	TOP5 volume \$25.97B , +235% w/w	pairwise corr 0.81 (was 0.52)

KEY FINDING · OBSERVATION (one weekly window)

Herding eased to 97.1% and, for the first time in the series, leaving the crowd was the mistake: agree-hit 54.8% against 37.6% on disagreement, with the margin curve taking its third shape in three weeks.

TL;DR

- **OBSERVATION -- Herding eased, and this week it paid.** Of 4,357 LOO-scoreable directional calls, 4,232 (97.1%) sided with peers' majority (issue #2: 99.8%); agree-hit was 54.8% [53.3%, 56.2%] vs. 37.6% [29.6%, 46.3%] for the 125 disagreeers, a pooled lift of +17.1pp. The disagree bucket is 20x issue #2's (n=6) and the two intervals no longer overlap. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a third shape.** 3 peers 56.4% (was 50.2%), 4 peers 48.9% (was 39.4%), 5 peers 54.4% (was 41.0%), full unanimity 56.6% (was 43.3%). Only the 4-peer tier sits below the 54.1% base; three weeks, three different orderings.
- **Daily-horizon unanimity flipped outright.** 1d/1d margin-6 hit 81.0% (n=84) and 1d/4h margin-6 hit 61.9% (n=84) -- issue #2 had the 1d/4h near-unanimous bucket at 0 for 28.
- **Herd failures did not disappear.** 437 cells had 6+ directional models on one side; the 5 worst were all-7-long swarms at ~68-69 mean confidence that resolved 0.0-14.3%, four of them inside Aug 18-19 as the tape turned.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #3 n: 468 / 700 / 1,128 / 1,855; field base this week 54.1% (issue #2: 43.8%). Every tier rose with the base rate; only the 4-peer tier finished below it. A 2-peer bucket exists this week too (n=81, 59.3%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #2.

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. 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	4,232	54.8%	[53.3%, 56.2%]	+17.1 (vs. disagree)
Disagree with LOO peers	125	37.6%	[29.6%, 46.3%]	ref.

Read with context. This is the series' first readable disagree bucket -- n=125 against 6 in issue #2 and 19 in issue #1 -- and the first time the two Wilson intervals do not overlap. It is still a single week of dependent observations in a regime that flipped: not evidence that herding is safe 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, LOO disagreement finally appears in every cell (issue #2 had all 6 disagreements in the hourly cell), and the agree-side lift is positive in all five.

FH / TF	Agree n	Agree hit	Dis. n	Dis. hit	Lift, pp	Excluded
1h / 1h	2,554	49.3%	51	45.1%	+4.2	484
4h / 4h	657	63.2%	50	36.0%	+27.2	133
4h / 1h	703	58.8%	18	33.3%	+25.4	132
1d / 1d	159	75.5%	4	0.0%	+75.5	8
1d / 4h	159	69.2%	2	0.0%	+69.2	16

"Excluded" = thin (fewer than 3 directional peers) or tied peer splits in that cell; not folded into agree or disagree. "Dis." = disagree. The two daily cells' disagree buckets (n=4 and n=2, both 0.0%) 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 models were active and LOO always drops one). The week-over-week chart is on page 1; the full per-tier numbers: 3 peers 56.4% (n=468), 4 peers 48.9% (n=700), 5 peers 54.4% (n=1,128), full unanimity 56.6% (n=1,855). Issue #2 asked whether the margin-curve inversion would persist: it did not. Last week's leader (3 peers, 50.2%) is still above base but no longer the top tier; unanimity went from 43.3% -- below base -- to 56.6%, the best tier of the week. Three issues in, the margin curve has taken three different shapes, which is itself the finding: no tier carries a stable meaning yet.

Daily-horizon herds: the flip

Issue #2's ugliest cells were the daily ones -- 1d/1d margin-5/6 at 19.4%/17.1% and 1d/4h near-unanimity 0 for 28. This week the same cells are the strongest in the report: 1d/1d unanimity 81.0% (n=84), 1d/4h unanimity 61.9% (n=84), and the whole 1d/1d agree bucket at 75.5% (n=159). The daily horizon did not become readable because the herd changed; it became readable because the mean absolute 1d move went from 0.73% to 4.28%. See the market check below.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 437 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side -- up from 299 in issue #2, in a week when the field made 40% more directional calls.

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Aug 18, 16:01	BTC	4h / 1h	long	7	69.3	0.0%
Aug 23, 22:01	BTC	1h / 1h	long	7	68.4	0.0%
Aug 19, 09:01	SOL	1h / 1h	long	7	68.3	0.0%
Aug 18, 19:01	SOL	1h / 1h	long	7	68.0	0.0%
Aug 18, 16:01	SOL	1h / 1h	long	7	68.0	14.3%

All 5 were long swarms at ~68-69 mean confidence; four resolved 0.0% and the fifth 14.3%. Four of the five landed inside Aug 18-19. 'Sideways' fell to 44.6% of mature forecasts (4,130 of 9,260) from 60.6% in issue #2 -- the field committed to a direction far more often this week, which is why both the directional pool and the unanimous-cell count grew.

Per-model: agree vs. disagree

Disagreement stopped being an anecdote. Per-model disagree counts run 2 to 53 this week (issue #2: 0 to 4), and three models -- gemini-3.1-pro (53), grok-4.5 (33) and deepseek-v4-pro (22) -- clear the N=10 floor. All three under-performed their own agree bucket.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	666	55.6%	5	40.0%
claude-opus-5	604	58.8%	3	33.3%
deepseek-v4-pro	621	52.3%	22	45.5%
gemini-3.1-pro	578	55.0%	53	37.7%
gpt-5.6-sol	600	54.5%	7	28.6%
grok-4.5	559	54.6%	33	36.4%
qwen-3.8-max	604	52.5%	2	0.0%

Individual per-model contrarian records are anecdotes, not findings: at n=2 to n=53 a handful of flipped calls swings a 'lift' by tens of points, so models are never ranked by contrarian lift here.

Market check: the tape woke up

OBSERVATION — the tape sped up and accuracy followed: mean |1d move| 0.73% -> 4.28% (483% rel), BTC realized vol 19.3% -> 58.7% (ann., hourly); field directional accuracy 43.8% -> 54.1% (+10.3pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -\$527 -> +\$889 (adjacent calendar weeks Aug 10-16 vs Aug 17-23; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 40.1% -> 53.5% — consistent with the trade-based hit rule.

Measure	Aug 10-16 (week A)	Aug 17-23 (week B)	Change
Mean 1d move (5 assets)	0.73%	4.28%	+483% rel
BTC realized vol (ann., hourly)	19.3%	58.7%	+39.4pp
Field directional accuracy	43.8%	54.1%	+10.3pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	40.1%	53.5%	+13.4pp
Field sim win-rate	29.7%	46.3%	+16.6pp
Field sim net PnL	-\$527.07	+\$889.37	--

Week A is exactly the issue-#2 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 71.1% is the audited daily-candle figure -- different estimators, both reported as measured.

Practical implications

- Three weeks, three margin-curve shapes. No consensus tier has shown a stable meaning, so a consensus filter still should not carry weight in a config.
- The agree-vs-disagree lift finally rests on a readable disagree bucket (n=125) and it points the opposite way from issue #2. Two opposite signs in two weeks is what a regime-dependent description looks like; treat +17.1pp as this week's weather.
- The daily-horizon flip (0 for 28 -> 81.0% on 1d/1d unanimity) tracks the market waking up, not a change in how the herd forms. Read the two together, never the consensus number alone.

Limitations

- Single week (Aug 17-23, 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 125 disagreeing calls; per-model counts run 2-53 and two cell-level disagree buckets are N<10. 773 calls (15.1% of the 5,130-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window ran 5 trend days of 7 (issue #2: 1 of 7); every comparison with issue #2 crosses two different regimes. Series density and lineage notes (wave 3): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 17-23) daily coverage is PARTIAL by design: the 1w series has the Mon Aug 17 anchor plus daily slots on Aug 22-23 only, and the 1M series has a daily slot on Aug 23 only; both sit outside this report's FH gate (1h/4h/1d) and appear only in the exclusions counter (1w 209, 1M 68). (2) After the report window — from Aug 24, 2026 — the grok line runs Grok 4.6; every grok forecast in this window and in the week-2 comparison is grok-4.5. (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 this issue, after an audit found the raw table mixes two exchanges; issue #2 row is as published.
- Research-to-date counter pinned from this issue: directional forecasts resolved inside the FH gate (1h/4h/1d) since Jul 11, counted at the issue cutoff (Aug 24 16:00 UTC). Issue #2 printed 23,458 under an earlier, unpinned definition; treat cross-issue counter deltas across the pin as definitional, not additive.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	9,590	Mature (scored pool)	9,260
OK in gate	9,260	-- of them directional	5,130
Out of gate (1w / 1M)	209 / 68	-- of them sideways	4,130
Invalid	53	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-08-17.json	Report cutoff	Mon Aug 24, 2026, 16:00 UTC
Generated at (pipeline)	2026-08-26 05:22 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,357 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Aug 24, 2026 cutoff): 33,809 directional forecasts resolved since Jul 11 · 11 models tracked (7 current + 4 archived legacy) · 5 assets · 5 horizons · hourly · 11 published reports.

Issue #3. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #2's open question -- does the margin-curve inversion persist, or is tier ordering week-to-week noise? -- closed as 'noise so far': the curve took a third, different shape. The new result is the sign flip on the pooled lift, on the first readable disagree bucket of the series. After the report window — from Aug 24, 2026 — the grok line runs Grok 4.6; every grok forecast in this window and in the week-2 comparison is grok-4.5. Engine 1.1 has powered the sandbox since Aug 18, i.e. from day 2 of this window; no cross-engine PnL comparisons are claimed.

What we're testing next

- Next issue: does the agree-side lift survive a fourth week -- and does it survive a flat week, or is it a trend-regime artifact?
- Does the 4-peer dip repeat, or is it this week's noise in the only below-base tier?
- Monthly test (Sep 2): the agreement curve across regimes at monthly n.

Related research

Report	Direct PDF link
Weekly Calibration #3	https://marketmania.ai/research/reports/weekly-calibration-2026-08-17.pdf
Weekly Model Watch #3	https://marketmania.ai/research/reports/model-watch-2026-08-17.pdf
Consensus Watch #2	https://marketmania.ai/research/reports/consensus-watch-2026-08-10.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_2026w34,
  title = {Consensus Watch #3: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year = {2026}, month = {August}, day = {26},
  url = {https://marketmania.ai/research/reports/consensus-watch-2026-08-17.pdf},
  note = {Methodology v1.1; window Aug 17-23, 2026 UTC; source weekly_metrics_2026-08-17.json}
}
```