

# Consensus Watch #6

September 15, 2026 · MarketMania Research · Weekly series · window Sep 7-13, 2026 UTC

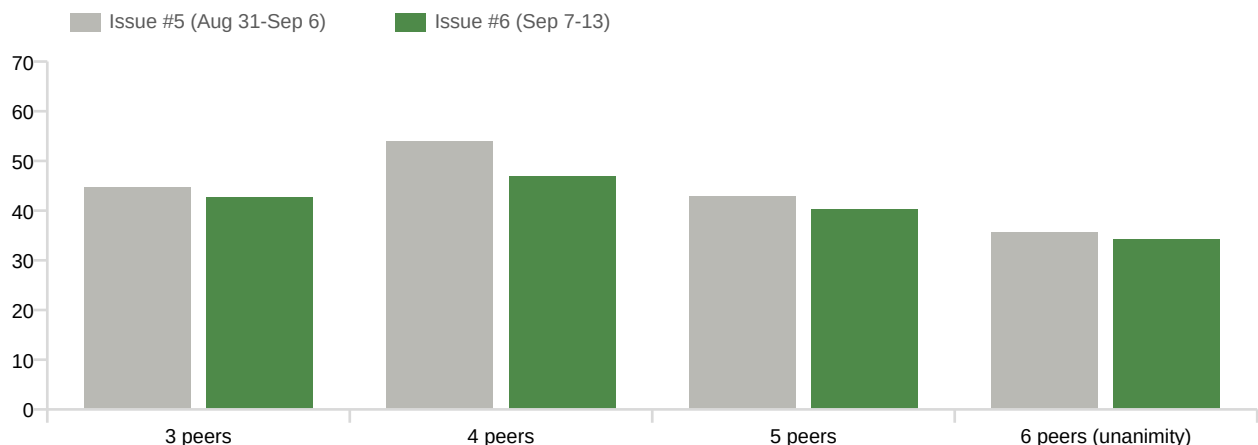
| RESEARCH SNAPSHOT                                               |                                  |                                          |                                           |
|-----------------------------------------------------------------|----------------------------------|------------------------------------------|-------------------------------------------|
| 3,788 LOO-scoreable calls                                       | 7 models (stable lineup)         | 680 thin/tied excluded                   | window <b>Sep 7-13, 2026 UTC</b>          |
| method = <b>leave-one-out</b><br>majority, 3+ directional peers | FH gate <b>1h / 4h / 1d</b>      | base field hit <b>39.6%</b>              | cutoff <b>Mon Sep 14, 2026, 16:00 UTC</b> |
| market: BTC net <b>-4.36%</b><br>(prior +3.42%)                 | ann vol <b>19.7%</b> (was 43.4%) | TOP5 volume <b>\$16.05B</b> , -1%<br>w/w | pairwise corr <b>0.74</b> (was 0.78)      |

KEY FINDING · OBSERVATION (one weekly window)

**Herding ran at 99.3% of leave-one-out scoreable calls and the pooled lift came in at -30.4pp: agree-hit 38.8% 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 3,788 LOO-scoreable directional calls, 3,762 (99.3%) sided with peers' majority (issue #5: 99.3%); agree-hit was 38.8% [37.2%, 40.4%] vs. 69.2% [50.0%, 83.5%] for the 26 disagreeers, a pooled lift of -30.4pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a sixth reading.** 3 peers 42.8% (was 44.7%), 4 peers 46.9% (was 54.0%), 5 peers 40.4% (was 43.0%), full unanimity 34.2% (was 35.7%). 3 of the 4 tiers sit above the 39.6% field base, and 4 of the 4 kept the side of the base they took in issue #5.
- **Daily-horizon unanimity fell below the field base.** 1d/1d margin-6 hit 0.0% (n=49) and 1d/4h margin-6 hit 14.3% (n=49); issue #5 printed 12.5% and 21.4% on the same two cells.
- **Herd failures did not disappear.** 417 cells had 6+ directional models on one side (issue #5: 392); the 5 worst sat at 67.4-68.1 mean confidence and resolved 0.0%-0.0%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #6 n: 400 / 550 / 984 / 1,813; field base this week 39.6% (issue #5: 42.8%). A 2-peer bucket exists this week too (n=15, 80.0%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #5.

## 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 two 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 first pure-4.6 against pure-4.6 comparison of the series.

## Pooled result: agree vs. disagree with the LOO consensus

| Group                   | n     | Hit rate | Wilson 95% CI  | Lift, pp             |
|-------------------------|-------|----------|----------------|----------------------|
| Agree with LOO peers    | 3,762 | 38.8%    | [37.2%, 40.4%] | -30.4 (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 26 in issue #5 and 36 in issue #4; 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 | 2,301   | 43.1%     | 22     | 68.2%    | -25.1    | 439      |
| 4h / 4h | 584     | 33.7%     | 3      | 66.7%    | -32.9    | 117      |
| 4h / 1h | 668     | 34.6%     | 1      | 100.0%   | -65.4    | 80       |
| 1d / 1d | 105     | 14.3%     | 0      | n/a      | n/a      | 27       |
| 1d / 4h | 104     | 24.0%     | 0      | n/a      | n/a      | 17       |

"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 42.8% (n=400), 4 peers 46.9% (n=550), 5 peers 40.4% (n=984), 6 peers 34.2% (n=1,813). Against the 39.6% field base, 3 tier(s) sit above and 1 below, and 4 of the 4 tiers kept the side of the base they took in issue #5. Six issues in, the margin curve has taken six readings.

## Daily-horizon herds: the 1d/1d unanimity cell went 0 for 49

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 0.0% (n=49) and 1d/4h unanimity 14.3% (n=49), both below the 39.6% field base; the whole 1d/1d agree bucket ran 14.3% (n=105) -- not the highest of the five FH/TF cells -- while 1d/4h ran 24.0% (n=104), the thinnest agree bucket in the table above. Issue #5 printed 12.5% and 21.4% on the same two unanimity cells. Both daily buckets are an order of magnitude thinner than the hourly one (n=2,301), so they are reported, not ranked.

## Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 417 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #5: 392).

| Slot (UTC)    | Symbol | FH / TF | Side  | Models | Mean conf | Hit rate |
|---------------|--------|---------|-------|--------|-----------|----------|
| Sep 12, 11:01 | BNB    | 1h / 1h | long  | 7      | 68.1      | 0.0%     |
| Sep 8, 16:01  | XRP    | 4h / 1h | long  | 7      | 67.9      | 0.0%     |
| Sep 12, 14:01 | ETH    | 1h / 1h | long  | 7      | 67.7      | 0.0%     |
| Sep 8, 07:01  | BTC    | 1h / 1h | short | 7      | 67.4      | 0.0%     |
| Sep 8, 16:01  | XRP    | 1h / 1h | long  | 7      | 67.4      | 0.0%     |

The 5 worst cells were long/short swarms at 67.4-68.1 mean confidence and resolved 0.0%-0.0%. 'Sideways' was 52.0% of mature forecasts (4,833 of 9,301) this week.

## Per-model: agree vs. disagree

Per-model disagree counts run 1 to 12 this week (issue #5: 0 to 11). 1 of the 7 lines clear the N=10 floor (gemini-3.1-pro 12); every other per-model contrarian cell below is insufficient.

| Model           | Agree n | Agree hit | Disagree n | Disagree hit |
|-----------------|---------|-----------|------------|--------------|
| claude-fable-5  | 546     | 36.8%     | 3          | 33.3%        |
| claude-opus-5   | 474     | 36.1%     | 1          | 100.0%       |
| deepseek-v4-pro | 565     | 38.8%     | 7          | 71.4%        |
| gemini-3.1-pro  | 513     | 41.3%     | 12         | 83.3%        |
| gpt-5.6-sol     | 594     | 40.4%     | 1          | 100.0%       |
| grok-4.6        | 490     | 39.4%     | 1          | 0.0%         |
| qwen-3.8-max    | 580     | 38.5%     | 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. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits two 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 first pure-4.6 against pure-4.6 comparison of the series.

## Market check: a down week on a quieter tape

OBSERVATION — mean |1d move| 2.10% -> 1.35% (-36% rel), BTC realized vol 35.2% -> 31.4% (ann., hourly); field directional accuracy 42.8% -> 39.6% (-3.3pp), 1 of 7 models improved, sim win-rate up for 1 of 7, field sim PnL -\$669.03 -> -\$1018.93 (adjacent calendar weeks Aug 31-Sep 6 vs Sep 7-13; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 51.6% -> 38.7% — the same direction as the trade-based hit rule.

| Measure                         | Aug 31-Sep 6 (week A) | Sep 7-13 (week B) | Change   |
|---------------------------------|-----------------------|-------------------|----------|
| Mean  1d move  (5 assets)       | 2.10%                 | 1.35%             | -36% rel |
| BTC realized vol (ann., hourly) | 35.2%                 | 31.4%             | -3.8pp   |
| Field directional accuracy      | 42.8%                 | 39.6%             | -3.3pp   |
| Models improving hit-rate       | --                    | 1 of 7            | --       |
| Raw price-sign accuracy         | 51.6%                 | 38.7%             | -12.8pp  |
| Field sim win-rate              | 35.6%                 | 32.1%             | -3.5pp   |
| Field sim net PnL               | -\$669.03             | -\$1018.93        | --       |

Week A is exactly the issue-#5 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, two windows back, inside issue #4's window). Week A (2026-08-31..2026-09-06) is pure grok-4.6 (it reproduces the published issue #5); week B is pure grok-4.6.

## Practical implications

- Six issues, six margin-curve readings. The tier ordering held for the first time -- 4 peers on top, unanimity at the bottom, as in issue #5 -- but all four tiers printed a lower rate than they did then (4 peers 54.0% -> 46.9%, full unanimity 35.7% -> 34.2%), so a consensus filter still should not carry weight in a config.
- The pooled lift is -30.4pp this week against -27.6pp in issue #5 and -6.0pp in issue #4. Three issues on the same side of zero, 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 -3.3pp week over week; 0 of the 4 consensus tiers moved up and 4 down against issue #5 (3 peers -2.0pp, 4 peers -7.1pp, 5 peers -2.7pp, unanimity -1.5pp).

## Limitations

- Single week (Sep 7-13, 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 1-12. 680 calls (15.2% of the 4,468-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window ran 2 trend days of 7 (issue #5: 3 of 7); every comparison with issue #5 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 6): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 7-13) daily coverage is FULL, as it was in issues #4 and #5: 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, two 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-#5 window), so every grok number in this issue and in issue #5 is a pure grok-4.6 line -- the grok week-over-week row is the first 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 #5 after the audit that found the raw table mixes two exchanges; issue #5 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 14 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4 and #5, which printed 42,905 at the Sep 7 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 47,373 is an additive step under one definition; counter deltas against issue #4 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 | 4,468                                     |
| Out of gate (1w / 1M)   | 490 / 490                      | -- of them sideways    | 4,833                                     |
| Invalid                 | 9                              | Pending (next issue)   | 0                                         |
| Late closes             | 0                              | Uptime, all grids      | 266 / 266 slots                           |
| Source file             | weekly_metrics_2026-09-07.json | Report cutoff          | Mon Sep 14, 2026, 16:00 UTC               |
| Generated at (pipeline) | 2026-09-15 15:24 UTC           | Methodology            | v1.1 (2026-08-10) · hash e66c7e8c864a2233 |

**THIS REPORT: 3,788 LOO-scored observations.** MARKETMANIA RESEARCH TO DATE (as of Sep 14, 2026 cutoff): 47,373 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 27 published reports.

**Issue #6.** Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #5 asked whether the pooled agree-vs-disagree sign keeps the same direction for a third issue running: it reads -30.4pp this week after -27.6pp in issue #5 (agree 38.8% vs disagree 69.2% on n=26), and whether the same margin tiers hold their side of the field base for a third issue: 4 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 and was pure grok-4.6 in issue #5; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits two 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 the pooled agree-vs-disagree sign keep the same direction for a fourth issue running?
- Do the same margin tiers hold their side of the field base for a fourth issue?
- Monthly series: the agreement curve across regimes at monthly n.

## Related research

| Report                | Direct PDF link                                                                                                                                                   |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Weekly Calibration #6 | <a href="https://marketmania.ai/research/reports/weekly-calibration-2026-09-07.pdf">https://marketmania.ai/research/reports/weekly-calibration-2026-09-07.pdf</a> |
| Weekly Model Watch #6 | <a href="https://marketmania.ai/research/reports/model-watch-2026-09-07.pdf">https://marketmania.ai/research/reports/model-watch-2026-09-07.pdf</a>               |
| Consensus Watch #5    | <a href="https://marketmania.ai/research/reports/consensus-watch-2026-08-31.pdf">https://marketmania.ai/research/reports/consensus-watch-2026-08-31.pdf</a>       |

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_2026w37,
  title = {Consensus Watch #6: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year  = {2026}, month = {September}, day = {15},
  url   = {https://marketmania.ai/research/reports/consensus-watch-2026-09-07.pdf},
  note  = {Methodology v1.1; window Sep 7-13, 2026 UTC; source weekly_metrics_2026-09-07.json}
}
```