# Consensus Watch #13

November 4, 2026 · MarketMania Research · Weekly series · window Oct 26-Nov 1, 2026 UTC

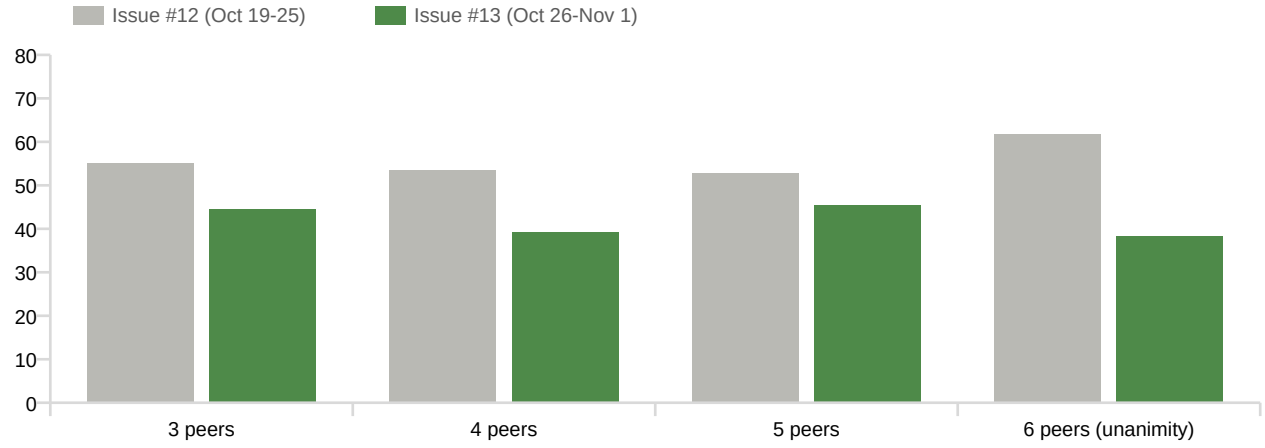
| RESEARCH SNAPSHOT | | | |
|---|---|---|---|
| **4,344** LOO-scoreable calls | **7** models (stable lineup) | **750** thin/tied excluded | window **Oct 26-Nov 1, 2026 UTC** |
| method = **leave-one-out** majority, 3+ directional peers | FH gate **1h / 4h / 1d** | base field hit **41.6%** | cutoff **Mon Nov 2, 2026, 16:00 UTC** |
| market: BTC net **-2.25%** (prior -2.39%) | ann vol **34.7%** (was 29.3%) | TOP5 volume **$14.86B**, +6% w/w | pairwise corr **0.93** (was 0.93) |

KEY FINDING · OBSERVATION (one weekly window)

**Herding ran at 98.7% of leave-one-out scoreable calls and the pooled lift came in at -8.5pp: agree-hit 40.6% against 49.1% on disagreement, on a disagree bucket of n=55.**

## TL;DR

- **OBSERVATION -- 98.7% of the field's leave-one-out calls ran with the herd.** Of 4,344 LOO-scoreable directional calls, 4,289 (98.7%) sided with peers' majority (issue #12: 98.9%); agree-hit was 40.6% [39.2%, 42.1%] vs. 49.1% [36.4%, 61.9%] for the 55 disagreers, a pooled lift of -8.5pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a thirteenth reading.** 3 peers 44.4% (was 55.0%), 4 peers 39.1% (was 53.5%), 5 peers 45.4% (was 52.8%), full unanimity 38.4% (was 61.7%). 2 of the 4 tiers sit above the 41.6% field base, and 1 of the 4 kept the side of the base they took in issue #12.
- **Daily-horizon unanimity stayed above the field base.** 1d/1d margin-6 hit 44.0% (n=84) and 1d/4h margin-6 hit 58.4% (n=77); issue #12 printed 100.0% and 94.5% on the same two cells.
- **Herd failures did not disappear.** 474 cells had 6+ directional models on one side (issue #12: 588); the 5 worst sat at 67.7-69.3 mean confidence and resolved 0.0%-0.0%.




Hit rate (%) by number of agreeing LOO peers, week over week. Issue #13 n: 392 / 655 / 942 / 2,261; field base this week 41.6% (issue #12: 57.0%). A 2-peer bucket exists this week too (n=39, 43.6%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #12.

## 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.

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

| Group | n | Hit rate | Wilson 95% CI | Lift, pp |
|---|---|---|---|---|
| Agree with LOO peers | 4,289 | 40.6% | [39.2%, 42.1%] | -8.5 (vs. disagree) |
| Disagree with LOO peers | 55 | 49.1% | [36.4%, 61.9%] | ref. |

Read with context. The disagree bucket is n=55 this week against 59 in issue #12 and 54 in issue #11; 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 | 2,578 | 43.8% | 46 | 52.2% | -8.4 | 473 |
| 4h / 4h | 715 | 35.9% | 3 | 66.7% | -30.7 | 130 |
| 4h / 1h | 719 | 33.2% | 4 | 25.0% | +8.2 | 106 |
| 1d / 1d | 145 | 35.9% | 1 | 0.0% | +35.9 | 17 |
| 1d / 4h | 132 | 50.0% | 1 | 0.0% | +50.0 | 24 |

"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 44.4% (n=392), 4 peers 39.1% (n=655), 5 peers 45.4% (n=942), 6 peers 38.4% (n=2,261). Against the 41.6% field base, 2 tier(s) sit above and 2 below, and 1 of the 4 tiers kept the side of the base they took in issue #12. Thirteen issues in, the margin curve has taken thirteen readings.

## Daily-horizon herds: unanimity above the field base

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 44.0% (n=84) and 1d/4h unanimity 58.4% (n=77), both above the 41.6% field base; the whole 1d/1d agree bucket ran 35.9% (n=145) -- not the highest of the five FH/TF cells -- while 1d/4h ran 50.0% (n=132), the thinnest agree bucket in the table above. Issue #12 printed 100.0% and 94.5% on the same two unanimity cells. Both daily buckets are more than an order of magnitude thinner than the hourly one (n=2,578), so they are reported, not ranked.

## Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 474 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #12: 588).

| Slot (UTC) | Symbol | FH / TF | Side | Models | Mean conf | Hit rate |
|---|---|---|---|---|---|---|
| Oct 30, 20:01 | BTC | 4h / 4h | short | 7 | 69.3 | 0.0% |
| Oct 31, 00:01 | BTC | 1d / 4h | short | 6 | 68.2 | 0.0% |
| Oct 29, 16:01 | BTC | 1h / 1h | long | 7 | 68.1 | 0.0% |
| Oct 30, 20:01 | BNB | 4h / 4h | short | 6 | 67.8 | 0.0% |
| Oct 30, 14:01 | BTC | 1h / 1h | short | 7 | 67.7 | 0.0% |

The 5 worst cells were long/short swarms at 67.7-69.3 mean confidence and resolved 0.0%-0.0%. 'Sideways' was 45.2% of mature forecasts (4,210 of 9,304) this week.

## Per-model: agree vs. disagree

Per-model disagree counts run 3 to 16 this week (issue #12: 4 to 16). 2 of the 7 lines clear the N=10 floor (gemini-3.1-pro 16, grok-4.7 13); every other per-model contrarian cell below is insufficient.

| Model | Agree n | Agree hit | Disagree n | Disagree hit |
|---|---|---|---|---|
| claude-fable-5 | 551 | 39.8% | 4 | 25.0% |
| claude-opus-5.5 | 548 | 40.7% | 7 | 42.9% |
| deepseek-v4-pro | 626 | 40.9% | 7 | 42.9% |
| gemini-3.1-pro | 610 | 39.5% | 16 | 43.8% |
| gpt-6.1-sol | 666 | 41.3% | 5 | 20.0% |
| grok-4.7 | 670 | 40.6% | 13 | 76.9% |
| qwen-3.8-max | 618 | 41.6% | 3 | 66.7% |

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.

## Market check: bigger daily moves on higher volume

OBSERVATION — mean |1d move| 1.35% -> 1.63% (+20% rel), BTC realized vol 27.5% -> 37.0% (ann., hourly); field directional accuracy 57.0% -> 41.6% (-15.3pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -$43.00 -> -$869.65 (adjacent calendar weeks Oct 19-25 vs Oct 26-Nov 1; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 56.0% -> 42.8% — the same direction as the trade-based hit rule.

| Measure | Oct 19-25 (week A) | Oct 26-Nov 1 (week B) | Change |
|---|---|---|---|
| Mean \|1d move\| (5 assets) | 1.35% | 1.63% | +20% rel |
| BTC realized vol (ann., hourly) | 27.5% | 37.0% | +9.5pp |
| Field directional accuracy | 57.0% | 41.6% | -15.3pp |
| Models improving hit-rate | -- | 0 of 7 | -- |
| Raw price-sign accuracy | 56.0% | 42.8% | -13.2pp |
| Field sim win-rate | 46.6% | 33.6% | -13.0pp |
| Field sim net PnL | -$43.00 | -$869.65 | -- |

Week A is exactly the issue-#12 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 note: the model cutover (2026-10-17T09:01:00+00:00, day 6 of 7 of its Mon-Sun week) sits two windows back, inside the published issue-#11 window, where it spliced three lines -- gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5, grok-4.6 -> grok-4.7. Week A (2026-10-19..2026-10-25, the published issue-#12 window) and week B (2026-10-26..2026-11-01) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #12. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits nine windows back (inside issue #4's window).

## Practical implications

- Thirteen issues, thirteen margin-curve readings. The tier ordering moved again -- the 5-peer tier went from the bottom to the top and the 6-peer tier fell from the top to the bottom -- and all four tiers printed a lower rate than in issue #12 (4 peers 53.5% -> 39.1%, full unanimity 61.7% -> 38.4%), so a consensus filter still should not carry weight in a config.
- The pooled lift is -8.5pp this week against +7.0pp in issue #12 and +1.8pp in issue #11. The sign fell below zero after three issues above it: 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 -15.3pp week over week; 0 of the 4 consensus tiers moved up and 4 down against issue #12 (3 peers -10.6pp, 4 peers -14.5pp, 5 peers -7.3pp, unanimity -23.3pp).

## Limitations

- Single week (Oct 26-Nov 1, 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 55 disagreeing calls; per-model counts run 3-16. 750 calls (14.7% of the 5,094-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window had trend on 1 of 7 days (issue #12: 1 of 7); every comparison with issue #12 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 13): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 26-Nov 1) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11 and #12: 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 488). (2) This window is the second full week on the builds that took over at Oct 17, 2026 09:01 UTC: claude-opus-5.5, gpt-6.1-sol and grok-4.7 carry no rows from the old builds in either week, and all 7 lines carry 1,470 rows each in both weeks, so every week-over-week row compares two full weeks on the same builds. (3) No roll-out slot of the flip falls inside either week: the two hourly slots before it (Oct 17, 07:01 and 08:01 UTC) lie two windows back, so this window's invalid counter holds none of their rows. (4) Past lineage flips: the Oct 17, 2026 09:01 UTC flip sits two windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits nine windows back (inside the issue-#4 window), and 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 #12 after the audit that found the raw table mixes two exchanges; issue #12 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 (Nov 2 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11 and #12, which printed 77,801 at the Oct 26 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 82,895 is an additive step under one definition; counter deltas against issue #11 and earlier remain definitional.

## Counters & lineage

| Counter | Value | Counter | Value |
|---|---|---|---|
| Forecasts total | 10,290 | Mature (scored pool) | 9,304 |
| OK in gate | 9,304 | -- of them directional | 5,094 |
| Out of gate (1w / 1M) | 490 / 488 | -- of them sideways | 4,210 |
| Invalid | 8 | Pending (next issue) | 0 |
| Late closes | 0 | Uptime, all grids | 266 / 266 slots |
| Source file | weekly_metrics_2026-10-26.json | Report cutoff | Mon Nov 2, 2026, 16:00 UTC |
| Generated at (pipeline) | 2026-11-03 11:35 UTC | Methodology | v1.1 (2026-08-10) · hash e66c7e8c864a2233 |

**THIS REPORT: 4,344 LOO-scored observations.** MARKETMANIA RESEARCH TO DATE (as of Nov 2, 2026 cutoff): 82,895 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 63 published reports.

**Issue #13.** Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #12 asked "the pooled agree-vs-disagree lift stayed above zero at +7.0pp for a third issue -- does the positive sign hold for a fourth issue?" -- it reads -8.5pp this week after +7.0pp in issue #12 (agree 40.6% vs disagree 49.1% on n=55), so the positive sign did not hold for a fourth issue; and "The 4-peer tier stayed below the field base for a third issue -- does it stay below for a fourth issue?" -- 1 of the 4 tiers kept their issue-#12 side of the base, and the 4-peer tier came in below it. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the second full week on the Oct 17 builds; week-over-week rows compare two full weeks on the same builds. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits nine 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: the pooled agree-vs-disagree lift fell below zero at -8.5pp after three issues above it -- does the negative sign hold for a second issue?
- The 4-peer tier stayed below the field base for a fourth issue -- does it stay below for a fifth issue?
- Monthly series: the agreement curve across regimes at monthly n.

## Related research

| Report | Direct PDF link |
|---|---|
| Weekly Calibration #13 | https://marketmania.ai/research/reports/weekly-calibration-2026-10-26.pdf |
| Weekly Model Watch #13 | https://marketmania.ai/research/reports/model-watch-2026-10-26.pdf |
| Consensus Watch #12 | https://marketmania.ai/research/reports/consensus-watch-2026-10-19.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_2026w44,
  title  = {Consensus Watch #13: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {4},
  url    = {https://marketmania.ai/research/reports/consensus-watch-2026-10-26.pdf},
  note   = {Methodology v1.1; window Oct 26-Nov 1, 2026 UTC; source weekly_metrics_2026-10-26.json}
}
```