Consensus Watch #15

November 18, 2026 · MarketMania Research · Weekly series · window Nov 9-15, 2026 UTC

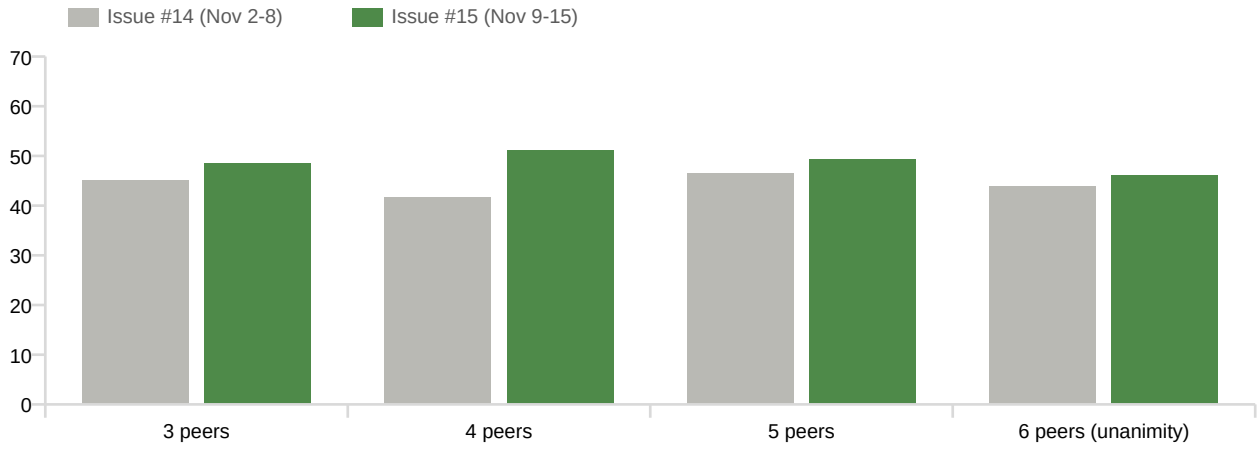
RESEARCH SNAPSHOT			
4,839 LOO-scoreable calls	**7** models (stable lineup)	**830** thin/tied excluded	window **Nov 9-15, 2026 UTC**
method = **leave-one-out** majority, 3+ directional peers	FH gate **1h / 4h / 1d**	base field hit **48.5%**	cutoff **Mon Nov 16, 2026, 16:00 UTC**
market: BTC net **-12.69%** (prior -4.47%)	ann vol **57.5%** (was 55.3%)	TOP5 volume **$18.77B**, +32% w/w	pairwise corr **0.95** (was 0.94)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 97.8% of leave-one-out scoreable calls and the pooled lift came in at -3.8pp: agree-hit 48.0% against 51.8% on disagreement, on a disagree bucket of n=108.

TL;DR

- **OBSERVATION -- 97.8% of the field's leave-one-out calls ran with the herd.** Of 4,839 LOO-scoreable directional calls, 4,731 (97.8%) sided with peers' majority (issue #14: 98.5%); agree-hit was 48.0% [46.6%, 49.5%] vs. 51.8% [42.5%, 61.1%] for the 108 disagreers, a pooled lift of -3.8pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a fifteenth reading.** 3 peers 48.6% (was 45.1%), 4 peers 51.1% (was 41.8%), 5 peers 49.4% (was 46.6%), full unanimity 46.1% (was 43.9%). 3 of the 4 tiers sit above the 48.5% field base, and 3 of the 4 kept the side of the base they took in issue #14.
- **Daily-horizon unanimity rose above the field base.** 1d/1d margin-6 hit 60.2% (n=98) and 1d/4h margin-6 hit 59.1% (n=105); issue #14 printed 36.4% and 37.5% on the same two cells.
- **Herd failures did not disappear.** 513 cells had 6+ directional models on one side (issue #14: 424); the 5 worst sat at 66.0-67.4 mean confidence and resolved 0.0%-0.0%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #15 n: 592 / 660 / 1,380 / 2,030; field base this week 48.5% (issue #14: 44.8%). A 2-peer bucket exists this week too (n=69, 43.5%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #14.

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	4,731	48.0%	[46.6%, 49.5%]	-3.8 (vs. disagree)
Disagree with LOO peers	108	51.8%	[42.5%, 61.1%]	ref.

Read with context. The disagree bucket is n=108 this week against 60 in issue #14 and 55 in issue #13; 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	2,647	48.8%	76	52.6%	-3.8	593
4h / 4h	923	44.5%	21	57.1%	-12.6	113
4h / 1h	755	39.1%	3	66.7%	-27.6	121
1d / 1d	201	69.2%	7	28.6%	+40.6	0
1d / 4h	205	65.8%	1	0.0%	+65.9	3

"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 48.6% (n=592), 4 peers 51.1% (n=660), 5 peers 49.4% (n=1,380), 6 peers 46.1% (n=2,030). Against the 48.5% field base, 3 tier(s) sit above and 1 below, and 3 of the 4 tiers kept the side of the base they took in issue #14. Fifteen issues in, the margin curve has taken fifteen readings.

Daily-horizon herds: unanimity above the field base

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 60.2% (n=98) and 1d/4h unanimity 59.1% (n=105), both above the 48.5% field base; the whole 1d/1d agree bucket ran 69.2% (n=201) -- the highest of the five FH/TF cells -- while 1d/4h ran 65.8% (n=205), the second-thinnest agree bucket in the table above. Issue #14 printed 36.4% and 37.5% on the same two unanimity cells. Both daily buckets are more than an order of magnitude thinner than the hourly one (n=2,647), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 513 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #14: 424).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Nov 10, 17:01	BTC	1h / 1h	short	7	67.4	0.0%
Nov 10, 16:01	BTC	4h / 4h	short	7	67.0	0.0%
Nov 10, 16:01	BTC	4h / 1h	short	7	66.7	0.0%
Nov 9, 08:01	SOL	4h / 4h	short	7	66.1	0.0%
Nov 9, 08:01	SOL	4h / 1h	short	7	66.0	0.0%

The 5 worst cells were short swarms at 66.0-67.4 mean confidence and resolved 0.0%-0.0%. 'Sideways' was 39.0% of mature forecasts (3,629 of 9,298) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 1 to 42 this week (issue #14: 1 to 22). 3 of the 7 lines clear the N=10 floor (gemini-3.1-pro 42, gpt-6.1-sol 20, grok-4.7 26); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	637	49.9%	4	25.0%
claude-opus-5.5	559	49.0%	8	12.5%
deepseek-v4-pro	717	46.4%	7	71.4%
gemini-3.1-pro	611	48.4%	42	54.8%
gpt-6.1-sol	782	49.5%	20	70.0%
grok-4.7	770	47.0%	26	46.2%
qwen-3.8-max	655	46.1%	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.

Market check: bigger daily moves on higher volume

OBSERVATION — mean |1d move| 2.16% -> 3.40% (+57% rel), BTC realized vol 41.0% -> 55.8% (ann., hourly); field directional accuracy 44.8% -> 48.5% (+3.7pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$464.89 -> +$160.79 (adjacent calendar weeks Nov 2-8 vs Nov 9-15; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 48.9% -> 54.4% — the same direction as the trade-based hit rule.

Measure	Nov 2-8 (week A)	Nov 9-15 (week B)	Change
Mean |1d move| (5 assets)	2.16%	3.40%	+57% rel
BTC realized vol (ann., hourly)	41.0%	55.8%	+14.8pp
Field directional accuracy	44.8%	48.5%	+3.7pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	48.9%	54.4%	+5.6pp
Field sim win-rate	37.4%	43.9%	+6.4pp
Field sim net PnL	-$464.89	+$160.79	--

Week A is exactly the issue-#14 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 four 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 (2026-11-02..2026-11-08, the published issue-#14 window) and week B (2026-11-09..2026-11-15) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #14. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits eleven windows back (inside issue #4's window).

Practical implications

- Fifteen issues, fifteen margin-curve readings. The tier ordering moved again -- the 4-peer tier rose from the bottom to the top, the 5-peer tier slipped to second and the 6-peer tier fell to the bottom -- and all four tiers printed a higher rate than in issue #14 (4 peers 41.8% -> 51.1%, full unanimity 43.9% -> 46.1%), so a consensus filter still should not carry weight in a config.
- The pooled lift is -3.8pp this week against -0.6pp in issue #14 and -8.5pp in issue #13. The sign stayed below zero for a third issue: 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 +3.7pp week over week; 4 of the 4 consensus tiers moved up and 0 down against issue #14 (3 peers +3.6pp, 4 peers +9.3pp, 5 peers +2.8pp, unanimity +2.2pp).

Limitations

- Single week (Nov 9-15, 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 108 disagreeing calls; per-model counts run 1-42. 830 calls (14.6% of the 5,669-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window had trend on 5 of 7 days (issue #14: 4 of 7); every comparison with issue #14 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 15): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Nov 9-15) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11, #12, #13 and #14: 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 489, 1M 487). (2) This window is the fourth 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 the 7 lines carry 1,470 rows in this window and 1,461 / 1,470 in week A, 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 four 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 four windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eleven 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 #14 after the audit that found the raw table mixes two exchanges; issue #14 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 16 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11, #12, #13 and #14, which printed 87,760 at the Nov 9 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 93,429 is an additive step under one definition; counter deltas against issue #13 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,298
OK in gate	9,298	-- of them directional	5,669
Out of gate (1w / 1M)	489 / 487	-- of them sideways	3,629
Invalid	16	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-11-09.json	Report cutoff	Mon Nov 16, 2026, 16:00 UTC
Generated at (pipeline)	2026-11-18 06:18 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

Provider note: grok-4.7 returned no valid answer for 13 of its 1,470 forecasts in this window (8 on Nov 9, 3 on Nov 10, 1 on Nov 11, 1 on Nov 15); those rows are counted in Invalid. The engine retries a call the provider rejects on input moderation without the news block (since Nov 6); 4 forecasts in this window (qwen-3.8-max: 1 on Nov 10, 2 on Nov 11, 1 on Nov 13) were produced that way and are flagged `news_stripped` in the open data. Hit rates in this report use the rows that resolved.

THIS REPORT: 4,839 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Nov 16, 2026 cutoff): 93,429 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 75 published reports.

Issue #15. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #14 asked "the pooled agree-vs-disagree lift stayed below zero at -0.6pp for a second issue -- does the negative sign hold for a third issue?" -- it reads -3.8pp this week after -0.6pp in issue #14 (agree 48.0% vs disagree 51.8% on n=108), so the negative sign held for a third issue; and "The 4-peer tier stayed below the field base for a fifth issue -- does it stay below for a sixth issue?" -- 3 of the 4 tiers kept their issue-#14 side of the base, and the 4-peer tier came in above it. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the fourth 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 eleven 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 stayed below zero at -3.8pp for a third issue -- does the negative sign hold for a fourth issue?
- The 4-peer tier moved above the field base after five issues below it -- does it stay above for a second issue?
- Monthly series: the agreement curve across regimes at monthly n.

Related research

Report	Direct PDF link
Weekly Calibration #15	https://marketmania.ai/research/reports/weekly-calibration-2026-11-09.pdf
Weekly Model Watch #15	https://marketmania.ai/research/reports/model-watch-2026-11-09.pdf
Consensus Watch #14	https://marketmania.ai/research/reports/consensus-watch-2026-11-02.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_2026w46,
  title  = {Consensus Watch #15: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {18},
  url    = {https://marketmania.ai/research/reports/consensus-watch-2026-11-09.pdf},
  note   = {Methodology v1.1; window Nov 9-15, 2026 UTC; source weekly_metrics_2026-11-09.json}
}
```