Consensus Watch #14

November 11, 2026 · MarketMania Research · Weekly series · window Nov 2-8, 2026 UTC

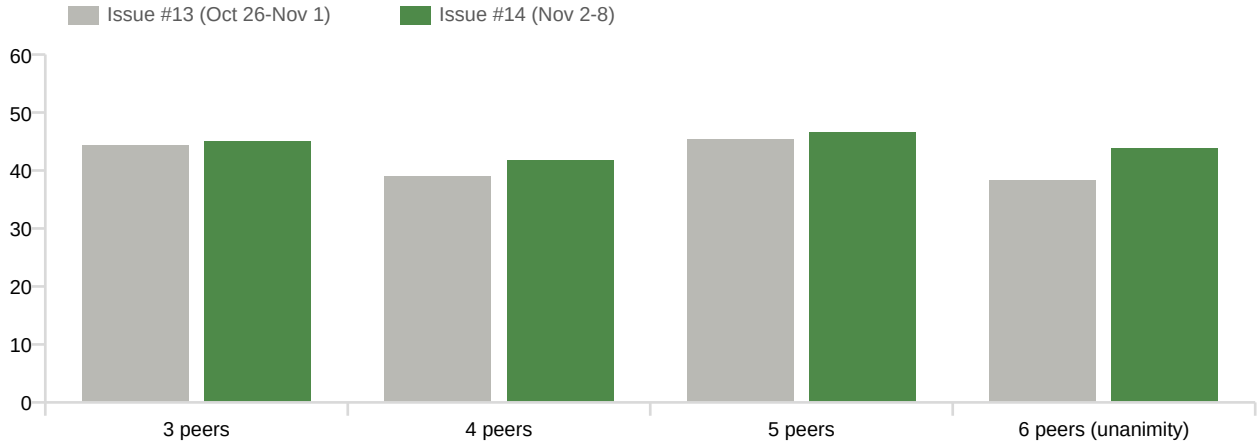
RESEARCH SNAPSHOT			
3,997 LOO-scoreable calls	**7** models (stable lineup)	**868** thin/tied excluded	window **Nov 2-8, 2026 UTC**
method = **leave-one-out** majority, 3+ directional peers	FH gate **1h / 4h / 1d**	base field hit **44.8%**	cutoff **Mon Nov 9, 2026, 16:00 UTC**
market: BTC net **-4.47%** (prior -2.25%)	ann vol **55.3%** (was 34.7%)	TOP5 volume **$14.26B**, -4% w/w	pairwise corr **0.94** (was 0.93)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 98.5% of leave-one-out scoreable calls and the pooled lift came in at -0.6pp: agree-hit 44.4% against 45.0% on disagreement, on a disagree bucket of n=60.

TL;DR

- **OBSERVATION -- 98.5% of the field's leave-one-out calls ran with the herd.** Of 3,997 LOO-scoreable directional calls, 3,937 (98.5%) sided with peers' majority (issue #13: 98.7%); agree-hit was 44.4% [42.9%, 46.0%] vs. 45.0% [33.1%, 57.5%] for the 60 disagreers, a pooled lift of -0.6pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a fourteenth reading.** 3 peers 45.1% (was 44.4%), 4 peers 41.8% (was 39.1%), 5 peers 46.6% (was 45.4%), full unanimity 43.9% (was 38.4%). 2 of the 4 tiers sit above the 44.8% field base, and 4 of the 4 kept the side of the base they took in issue #13.
- **Daily-horizon unanimity fell below the field base.** 1d/1d margin-6 hit 36.4% (n=77) and 1d/4h margin-6 hit 37.5% (n=112); issue #13 printed 44.0% and 58.4% on the same two cells.
- **Herd failures did not disappear.** 424 cells had 6+ directional models on one side (issue #13: 474); the 5 worst sat at 65.7-68.0 mean confidence and resolved 0.0%-28.6%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #14 n: 508 / 515 / 768 / 2,107; field base this week 44.8% (issue #13: 41.6%). A 2-peer bucket exists this week too (n=39, 56.4%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #13.

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	3,937	44.4%	[42.9%, 46.0%]	-0.6 (vs. disagree)
Disagree with LOO peers	60	45.0%	[33.1%, 57.5%]	ref.

Read with context. The disagree bucket is n=60 this week against 55 in issue #13 and 59 in issue #12; 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	2,266	43.5%	51	43.1%	+0.4	571
4h / 4h	770	48.6%	2	50.0%	-1.4	110
4h / 1h	604	40.9%	5	60.0%	-19.1	150
1d / 1d	140	50.7%	2	50.0%	+0.7	24
1d / 4h	157	44.6%	0	n/a	n/a	13

"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 45.1% (n=508), 4 peers 41.8% (n=515), 5 peers 46.6% (n=768), 6 peers 43.9% (n=2,107). Against the 44.8% field base, 2 tier(s) sit above and 2 below, and 4 of the 4 tiers kept the side of the base they took in issue #13. Fourteen issues in, the margin curve has taken fourteen readings.

Daily-horizon herds: unanimity below the field base

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 36.4% (n=77) and 1d/4h unanimity 37.5% (n=112), both below the 44.8% field base; the whole 1d/1d agree bucket ran 50.7% (n=140) -- the highest of the five FH/TF cells -- while 1d/4h ran 44.6% (n=157), the second-thinnest agree bucket in the table above. Issue #13 printed 44.0% and 58.4% on the same two unanimity cells. Both daily buckets are more than an order of magnitude thinner than the hourly one (n=2,266), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 424 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #13: 474).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Nov 5, 00:01	BTC	1d / 4h	long	7	68.0	0.0%
Nov 4, 16:01	BTC	4h / 4h	long	7	67.4	0.0%
Nov 3, 16:01	ETH	4h / 1h	long	7	67.1	0.0%
Nov 4, 09:01	ETH	1h / 1h	long	7	66.4	28.6%
Nov 4, 16:01	SOL	4h / 4h	long	7	65.7	0.0%

The 5 worst cells were long swarms at 65.7-68.0 mean confidence and resolved 0.0%-28.6%. 'Sideways' was 47.4% of mature forecasts (4,385 of 9,250) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 1 to 22 this week (issue #13: 3 to 16). 2 of the 7 lines clear the N=10 floor (gemini-3.1-pro 14, grok-4.7 22); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	515	43.5%	2	50.0%
claude-opus-5.5	485	45.1%	6	50.0%
deepseek-v4-pro	580	44.3%	7	57.1%
gemini-3.1-pro	570	43.5%	14	35.7%
gpt-6.1-sol	615	44.2%	8	37.5%
grok-4.7	609	45.3%	22	45.5%
qwen-3.8-max	563	44.8%	1	100.0%

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 lower volume

OBSERVATION — mean |1d move| 1.63% -> 2.16% (+33% rel), BTC realized vol 37.0% -> 41.0% (ann., hourly); field directional accuracy 41.6% -> 44.8% (+3.1pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$869.65 -> -$464.89 (adjacent calendar weeks Oct 26-Nov 1 vs Nov 2-8; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 42.8% -> 48.9% — the same direction as the trade-based hit rule.

Measure	Oct 26-Nov 1 (week A)	Nov 2-8 (week B)	Change
Mean |1d move| (5 assets)	1.63%	2.16%	+33% rel
BTC realized vol (ann., hourly)	37.0%	41.0%	+4.0pp
Field directional accuracy	41.6%	44.8%	+3.1pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	42.8%	48.9%	+6.0pp
Field sim win-rate	33.6%	37.4%	+3.8pp
Field sim net PnL	-$869.65	-$464.89	--

Week A is exactly the issue-#13 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 three 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 (2026-10-26..2026-11-01, the published issue-#13 window) and week B (2026-11-02..2026-11-08) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #13. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits ten windows back (inside issue #4's window).

Practical implications

- Fourteen issues, fourteen margin-curve readings. The tier ordering moved again -- the 5-peer tier held the top, the 6-peer tier rose from the bottom to third and the 4-peer tier fell to the bottom -- and all four tiers printed a higher rate than in issue #13 (4 peers 39.1% -> 41.8%, full unanimity 38.4% -> 43.9%), so a consensus filter still should not carry weight in a config.
- The pooled lift is -0.6pp this week against -8.5pp in issue #13 and +7.0pp in issue #12. The sign stayed below zero for a second 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 +3.1pp week over week; 4 of the 4 consensus tiers moved up and 0 down against issue #13 (3 peers +0.7pp, 4 peers +2.7pp, 5 peers +1.2pp, unanimity +5.5pp).

Limitations

- Single week (Nov 2-8, 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 60 disagreeing calls; per-model counts run 1-22. 868 calls (17.8% of the 4,865-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window had trend on 4 of 7 days (issue #13: 1 of 7); every comparison with issue #13 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 14): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Nov 2-8) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11, #12 and #13: 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 487). (2) This window is the third 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 1,461 / 1,470 rows in this window and 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 three 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 three windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits ten 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 #13 after the audit that found the raw table mixes two exchanges; issue #13 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 9 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11, #12 and #13, which printed 82,895 at the Nov 2 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 87,760 is an additive step under one definition; counter deltas against issue #12 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,281	Mature (scored pool)	9,250
OK in gate	9,250	-- of them directional	4,865
Out of gate (1w / 1M)	490 / 487	-- of them sideways	4,385
Invalid	54	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-11-02.json	Report cutoff	Mon Nov 9, 2026, 16:00 UTC
Generated at (pipeline)	2026-11-10 14:05 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

Provider incident: on Nov 5 the provider behind qwen-3.8-max rejected 46 of its 210 first-attempt forecasts with an input-moderation error; those rows are counted in Invalid. From Nov 6 the engine retries such a call without the news block; 2 forecasts in this window (qwen-3.8-max: 2 on Nov 7) 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: 3,997 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Nov 9, 2026 cutoff): 87,760 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 71 published reports.

Issue #14. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #13 asked "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?" -- it reads -0.6pp this week after -8.5pp in issue #13 (agree 44.4% vs disagree 45.0% on n=60), so the negative sign held for a second issue; and "The 4-peer tier stayed below the field base for a fourth issue -- does it stay below for a fifth issue?" -- 4 of the 4 tiers kept their issue-#13 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 third 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 ten 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 -0.6pp for a second issue -- does the negative sign hold for a third issue?
- The 4-peer tier stayed below the field base for a fifth issue -- does it stay below for a sixth issue?
- Monthly series: the agreement curve across regimes at monthly n.

Related research

Report	Direct PDF link
Weekly Calibration #14	https://marketmania.ai/research/reports/weekly-calibration-2026-11-02.pdf
Weekly Model Watch #14	https://marketmania.ai/research/reports/model-watch-2026-11-02.pdf
Consensus Watch #13	https://marketmania.ai/research/reports/consensus-watch-2026-10-26.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_2026w45,
  title  = {Consensus Watch #14: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {11},
  url    = {https://marketmania.ai/research/reports/consensus-watch-2026-11-02.pdf},
  note   = {Methodology v1.1; window Nov 2-8, 2026 UTC; source weekly_metrics_2026-11-02.json}
}
```