Consensus Watch #1

August 10, 2026 · MarketMania Research · Weekly series

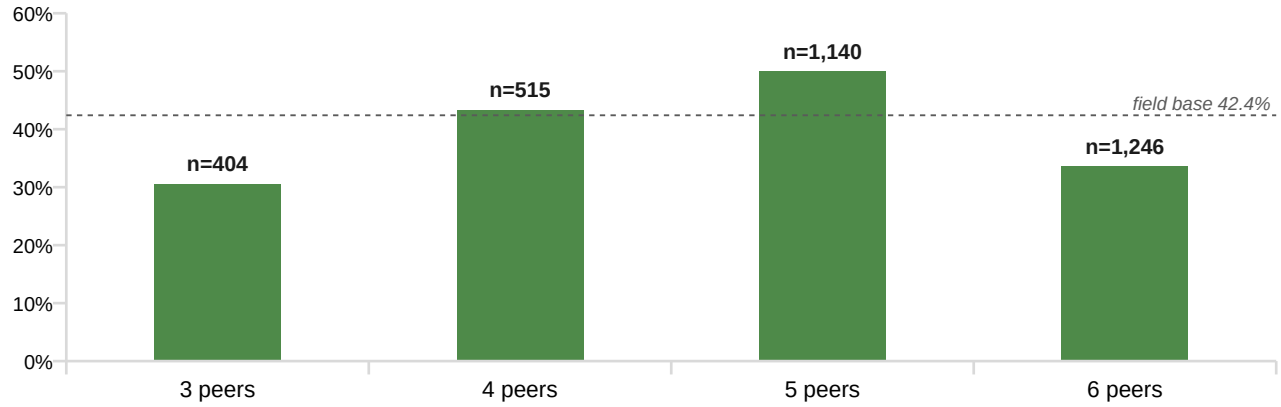
RESEARCH SNAPSHOT			
3,327 LOO-scoreable calls	**8** models (qwen-3.7 legacy to Aug 5)	**715** thin/tied excluded	window **Aug 3-9, 2026 UTC**
method = **leave-one-out** majority	3+ directional peers required	base field hit **42.4%**	cutoff **Mon Aug 10, 2026, 16:00 UTC**

KEY FINDING · OBSERVATION (one weekly window)

More agreement did not mean more reliability: full peer unanimity was the weakest high-consensus tier this week.

TL;DR

- OBSERVATION -- **Herding was near-total, and it did not pay.** Of 3,327 LOO-scoreable directional calls, 3,308 (99.4%) sided with peers' majority (the other 19 were all forecast-horizon 1h calls); agree-hit was 40.4% vs. 63.2% for those 19 disagreers, a pooled lift of -22.8pp on n=19 with a Wilson 95% CI of [41.0%, 80.9%] that overlaps the agree bucket's own CI [38.7%, 42.1%]. Descriptive, one week, not a claim of a durable edge.
- **The margin curve is not monotonic.** Hit rate ran 30.5% at 3 agreeing peers (n=404), up to 43.3% at 4 (n=515) and a peak of 49.9% at 5 (n=1,140), then back down to 33.6% once all peers agreed (6 of 6, n=1,246) — the largest bucket in the data. More agreement did not mean a safer call.
- **Herd failures were real.** 367 cells had 6 or more directional models on the same side; the week's 5 worst were total wipeouts — all 7 active models long, 0% hit — and all 5 were long calls in a week where 'sideways' was the field's single most common call (55.7% of mature forecasts).
- **Nobody escaped the herd.** Every model's LOO-disagree count this week is 0-6, far too small to rank models by contrarian lift; grok-4.5's 4 contrarian calls all hit and claude-fable-5's 2 both missed, but both are anecdotes, not findings.



This week's hit rate by number of agreeing LOO peers (dashed line = 42.4% field base rate). Full unanimity (6 of 6) was the weakest of the four plotted tiers -- full numbers and discussion in Margin, page 2.

Why it matters

MarketMania publishes forecasts from 8 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. This leave-one-out (LOO) design exists because the obvious alternative — comparing a model to a consensus that includes its own vote — is biased in the model's favor: with the model counted inside the index, its own vote nudges that index toward itself, mechanically inflating the agree bucket's measured hit rate and the resulting lift. 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

Across the full week and all forecast horizons: does a model's call agreeing with its LOO peers predict a better outcome than disagreeing?

Group	n	Hit rate	Wilson 95% CI	Lift, pp
Agree with LOO peers	3,308	40.4%	[38.7%, 42.1%]	**-22.8** (vs. disagree)
Disagree with LOO peers	19	63.2%	[41.0%, 80.9%]	ref.

Read with caution. The -22.8pp pooled lift above rests on just 19 disagreeing calls. Its Wilson 95% CI, [41.0%, 80.9%], overlaps the agree bucket's CI, [38.7%, 42.1%]. This is a descriptive, single-week result — not evidence that contrarian calls are more reliable, and not a claim that will necessarily hold up next week.

By forecast horizon / timeframe

The pooled result above mixes 5 different FH/TF cells. Broken out, LOO disagreement only happened in the FH 1h / TF 1h cell this week (19 calls, all of them); the other four cells had zero LOO-disagreeing directional calls, so no lift is defined for them — every model that had a valid majority there simply went along with it.

FH / TF	Agree n	Agree hit	Dis. n	Dis. hit	Lift, pp	Excluded
1h / 1h	2,070	41.7%	**19**	**63.2%**	-21.4	472
4h / 4h	562	39.2%	0	n/a	n/a	103
4h / 1h	544	37.7%	0	n/a	n/a	103
1d / 4h	77	33.8%	0	n/a	n/a	14
1d / 1d	55	38.2%	0	n/a	n/a	23

"Excluded" = thin (fewer than 3 directional peers) or tied peer splits in that cell; not folded into agree or disagree. "Dis." = disagree.

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 (as many as 6 of 6, since at most 7 models were active in any slot this week and LOO always drops one). The chart plotting this week's hit rate against that peer-agreement count appears on page 1, right after the TL;DR; the full per-tier numbers are below.

Hit rate by number of agreeing LOO peers, this week only: 3 peers 30.5% (n=404), 4 peers 43.3% (n=515), 5 peers 49.9% (n=1,140), 6 peers / full unanimity 33.6% (n=1,246). Only the 4- and 5-peer tiers beat the week's 42.4% field base rate (dashed line); the thinnest-majority and the fully unanimous tiers both fell below it. A 2-peer bucket also exists (n=3, 100% hit) but is far too small to plot or interpret — footnote only, not part of the trend.

The curve rises through 3, 4 and 5 agreeing peers, then drops sharply at full unanimity: going from 5-peer to 6-peer (full) agreement cost 16.3pp of hit rate this week, landing barely above the thinnest-majority bucket. Among the three higher-agreement tiers (4, 5 and 6 peers), full unanimity was the worst performer — stronger consensus did not mean a safer call.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 367 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side — effectively the whole active peer set agreeing before the outcome was known. The 5 worst of those herds were total wipeouts: every one of the 7 active models went the same direction, and the hit rate was zero.

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Aug 4, 19:01	BTC	1h / 1h	long	7	71.7	**0.0%**
Aug 3, 16:01	XRP	4h / 1h	long	7	70.0	0.0%
Aug 5, 20:01	BTC	4h / 1h	long	7	69.9	0.0%
Aug 5, 06:01	BNB	1h / 1h	long	7	69.7	0.0%
Aug 9, 16:01	BNB	4h / 4h	long	7	69.7	0.0%

All 5 worst herds were long calls, in a week where the field's own forecast mix leaned toward 'sideways' more than half the time (5,079 of 9,121 mature forecasts, 55.7%) — a 7-model long swarm was already a directional outlier before it missed. The worst of the five, the Aug 4, 19:01 UTC BTC 1h/1h slot (7 models long, mean confidence 71.7), was also individually the week's highest-confidence single miss: qwen-3.7-max went long on that same slot at 85% confidence, and the position still expired against it.

Per-model: agree vs. disagree

No model was immune to the herd. Every model's count of LOO-disagreeing calls this week is small — 0 to 6 out of several hundred agree calls each — which is exactly why models are not ranked by individual agree/disagree lift here; at these n, a single flipped call swings a 'lift' by tens of points.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	526	40.5%	2	**0.0%**
claude-opus-5	438	40.2%	0	n/a
deepseek-v4-pro	311	40.2%	3	66.7%
gemini-3.1-pro	494	39.3%	6	66.7%
gpt-5.6-sol	527	40.4%	0	n/a
grok-4.5	489	41.3%	4	**100.0%**
qwen-3.7-max	168	39.9%	1	100.0%
qwen-3.8-max	355	41.1%	3	33.3%

Lift is not rankable at these n. Two rows stand out only as descriptive anecdotes: grok-4.5's 4 contrarian calls all hit (4/4), and claude-fable-5's 2 contrarian calls both missed (0/2). claude-opus-5 and gpt-5.6-sol did not disagree with their LOO peers even once this week (n=0). None of this supports ranking models by 'contrarian skill' — the counts are too small to mean anything beyond this week's anecdote.

Smart Consensus

Deferred this issue. Quoted verbatim from this week's metrics: "Smart-vs-Outlook deferred: as-of weights pipeline not yet recording; will appear once weights are logged per-slot (methodology 3.2.3 fallback)" This issue's consensus is therefore an equal-weight peer majority only; a confidence- or performance-weighted 'Smart Consensus' comparison will appear in a future issue once that pipeline is recording.

Practical implications

- This week, peer-unanimity behaved like a crowding signal, not a safety signal — the largest 'everyone agrees' bucket (6 peers, n=1,246) hit worse (33.6%) than the 4- and 5-peer tiers below it. Treat a fully unanimous herd as a prompt to check position sizing, not as confirmation.
- Treat any consensus filter as regime-dependent, not as a fixed rule. This week's -22.8pp 'agree vs. disagree' lift (n=19 disagree; Wilson CI [41.0%, 80.9%], overlapping the agree bucket's CI) should not be read as 'contrarian calls are better' — it is one thin week, reported descriptively, not a strategy.
- Consensus Watch is a living series: each issue adds one more week to the LOO comparison and one more point on the margin curve. Only after several issues will it be possible to ask whether this week's non-monotonic curve and negative pooled lift are a pattern or noise.

Limitations

- Single week (Aug 3-9, 2026 UTC). Every finding in this issue is descriptive for this window only; no claim is made about next week or about any model's underlying skill.
- Observations are not independent: the same models watch overlapping symbol / FH / TF cells hour after hour, so LOO agree/disagree calls made close in time or on the same slot are correlated, not i.i.d. draws. Wilson confidence intervals in this issue are descriptive, not inferential.
- The design is correlational, not causal. All models see the same market data and broadly similar priors, so 'agreement' and 'hit' can rise together simply because a slot was easy to read that week — shared inputs mean shared blind spots, and a readable market can lift agreement and hit rate together without either causing the other. LOO removes self-match bias, not this confound.
- The headline contrarian result rests on 19 disagreeing calls pooled across all models, all of them in FH 1h; per-model disagree counts run 0-6, too small to rank models by contrarian lift.
- 715 calls (17.7% of the 4,042-call mature directional universe) were thin (fewer than 3 directional peers in the cell) or tied (peers split evenly) and are excluded from every consensus table in this issue — they are not folded into agree or disagree either way.

Counters & lineage

Metric	Value
Forecasts total	9,194
OK in gate	9,121
Out of gate (weekly window)	70
Invalid	3
Mature (resolved by cutoff)	9,121
Mature, directional	4,042
Mature, sideways	5,079
Pending to next issue	0
Late closes	0
Unanimous herds (6+ peers, same side)	367
LOO thin/tied (excluded from consensus tables)	715

Source: weekly_metrics_2026-08-03.json, generated 2026-08-11 14:24 UTC. **Methodology:** v1.1 (2026-08-10), hash e66c7e8c864a2233. **Snapshot principle:** each Consensus Watch issue is a frozen snapshot as of its Monday 16:00 UTC cutoff (Mon Aug 10, 2026, 16:00 UTC) — forecasts that mature after the cutoff roll into next week's issue rather than being backfilled here (pending to next issue: 0; late closes: 0). Past issues are never restated when later data arrives; a correction, if ever needed, is published as a note in a future issue.

THIS REPORT: 3,327 LOO-scored observations
MARKETMANIA RESEARCH TO DATE (as of Aug 10, 2026 cutoff): 14,151 resolved forecasts since Jul 11, 2026 · 9 models tracked (7 current frontier + 2 archived legacy generations) · 5 assets · 5 forecast horizons · hourly cadence · 4 published research reports

NEW series — Issue #1. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data, and coverage of the slower forecast horizons (especially 1d, 78-91 scoreable subjects this week vs. 2,561 for 1h/1h) widens as more weeks mature. Read any single issue descriptively, and watch the margin curve across issues before drawing conclusions.

What we're testing next

- **Next issue:** does full unanimity continue to underperform?
- **Monthly test:** does the agreement curve survive across multiple weeks and regimes?

Related research

Stability Index Run 2 — marketmania.ai/research/reports/si-run-2.pdf
Weekly Calibration #1 — marketmania.ai/research/reports/weekly-calibration-2026-08-03.pdf
Weekly Model Watch #1 — marketmania.ai/research/reports/model-watch-2026-08-03.pdf

These three weekly reports — Consensus Watch, Weekly Calibration and Weekly Model Watch — publish together as one wave each week.

CITE THIS REPORT

```
@misc{mm_consensus_watch_2026w32,
  title  = {Consensus Watch #1: does agreeing with the model consensus make a forecast more reliable?},
  author = {{MarketMania Research}},
  year   = {2026}, month = {August}, day = {10},
  url    = {https://marketmania.ai/research},
  note   = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Aug 3-9, 2026 UTC; source
weekly_metrics_2026-08-03.json}
}
```