

Consensus Watch #8

September 30, 2026 · MarketMania Research · Weekly series · window Sep 21-27, 2026 UTC

RESEARCH SNAPSHOT

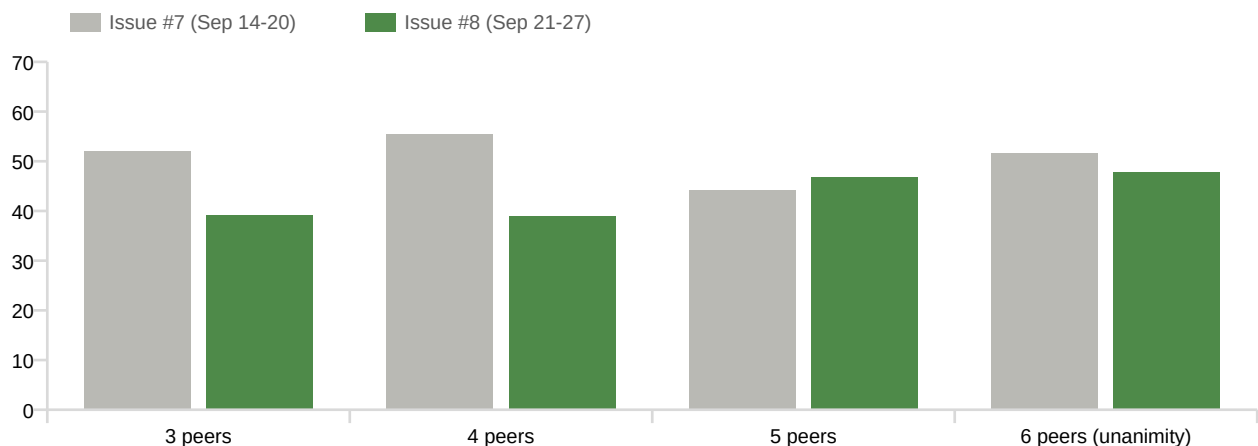
4,245 LOO-scoreable calls	7 models (stable lineup)	713 thin/tied excluded	window Sep 21-27, 2026 UTC
method = leave-one-out majority, 3+ directional peers	FH gate 1h / 4h / 1d	base field hit 46.8%	cutoff Mon Sep 28, 2026, 16:00 UTC
market: BTC net +4.06% (prior +5.64%)	ann vol 52.2% (was 51.0%)	TOP5 volume \$22.43B , +20% w/w	pairwise corr 0.71 (was 0.93)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 99.3% of leave-one-out scoreable calls and the pooled lift came in at -14.0pp: agree-hit 46.0% against 60.0% on disagreement, on a disagree bucket of n=30.

TL;DR

- **OBSERVATION -- 99.3% of the field's leave-one-out calls ran with the herd.** Of 4,245 LOO-scoreable directional calls, 4,215 (99.3%) sided with peers' majority (issue #7: 98.9%); agree-hit was 46.0% [44.5%, 47.5%] vs. 60.0% [42.3%, 75.4%] for the 30 disagreeers, a pooled lift of -14.0pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took an eighth reading.** 3 peers 39.1% (was 52.1%), 4 peers 39.0% (was 55.5%), 5 peers 46.8% (was 44.2%), full unanimity 47.8% (was 51.6%). 1 of the 4 tiers sit above the 46.8% field base, and 2 of the 4 kept the side of the base they took in issue #7.
- **Daily-horizon unanimity split around the field base.** 1d/1d margin-6 hit 51.7% (n=147) and 1d/4h margin-6 hit 39.6% (n=91); issue #7 printed 38.4% and 53.8% on the same two cells.
- **Herd failures did not disappear.** 505 cells had 6+ directional models on one side (issue #7: 539); the 5 worst sat at 69.0-70.4 mean confidence and resolved 0.0%-14.3%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #8 n: 304 / 485 / 876 / 2,541; field base this week 46.8% (issue #7: 50.7%). A 2-peer bucket exists this week too (n=9, 66.7%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #7.

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 four 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 third 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	4,215	46.0%	[44.5%, 47.5%]	-14.0 (vs. disagree)
Disagree with LOO peers	30	60.0%	[42.3%, 75.4%]	ref.

Read with context. The disagree bucket is n=30 this week against 52 in issue #7 and 26 in issue #6; 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	2,565	46.5%	21	57.1%	-10.6	460
4h / 4h	677	48.1%	2	50.0%	-1.8	94
4h / 1h	641	42.1%	3	100.0%	-57.9	138
1d / 1d	192	45.8%	4	50.0%	-4.2	10
1d / 4h	140	43.6%	0	n/a	n/a	11

"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 39.1% (n=304), 4 peers 39.0% (n=485), 5 peers 46.8% (n=876), 6 peers 47.8% (n=2,541). Against the 46.8% field base, 1 tier(s) sit above and 3 below, and 2 of the 4 tiers kept the side of the base they took in issue #7. Eight issues in, the margin curve has taken eight readings.

Daily-horizon herds: 1d/1d unanimity above the base, 1d/4h below

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 51.7% (n=147) and 1d/4h unanimity 39.6% (n=91), split around the 46.8% field base; the whole 1d/1d agree bucket ran 45.8% (n=192) -- not the highest of the five FH/TF cells -- while 1d/4h ran 43.6% (n=140), the thinnest agree bucket in the table above. Issue #7 printed 38.4% and 53.8% on the same two unanimity cells. Both daily buckets are an order of magnitude thinner than the hourly one (n=2,565), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 505 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #7: 539).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Sep 26, 00:01	SOL	1d / 1d	long	7	70.4	0.0%
Sep 22, 00:01	SOL	1d / 1d	long	7	70.4	0.0%
Sep 21, 01:01	SOL	1h / 1h	long	7	69.7	14.3%
Sep 21, 01:01	XRP	1h / 1h	long	7	69.1	0.0%
Sep 26, 16:01	SOL	4h / 4h	long	7	69.0	0.0%

The 5 worst cells were long swarms at 69.0-70.4 mean confidence and resolved 0.0%-14.3%. 'Sideways' was 46.5% of mature forecasts (4,315 of 9,273) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 0 to 16 this week (issue #7: 0 to 30). 1 of the 7 lines clear the N=10 floor (gemini-3.1-pro 16); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	620	47.3%	2	0.0%
claude-opus-5	585	47.5%	1	0.0%
deepseek-v4-pro	618	44.2%	5	40.0%
gemini-3.1-pro	625	48.3%	16	75.0%
gpt-5.6-sol	636	46.4%	3	66.7%
grok-4.6	495	44.9%	3	66.7%
qwen-3.8-max	636	43.2%	0	n/a

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 four 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 third pure-4.6 against pure-4.6 comparison of the series.

Market check: an up week with smaller daily moves

OBSERVATION — mean |1d move| 2.69% -> 1.83% (-32% rel), BTC realized vol 35.1% -> 35.8% (ann., hourly); field directional accuracy 50.7% -> 46.8% (-3.9pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -\$337.72 -> -\$315.68 (adjacent calendar weeks Sep 14-20 vs Sep 21-27; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 54.2% -> 49.2% — the same direction as the trade-based hit rule.

Measure	Sep 14-20 (week A)	Sep 21-27 (week B)	Change
Mean 1d move (5 assets)	2.69%	1.83%	-32% rel
BTC realized vol (ann., hourly)	35.1%	35.8%	+0.7pp
Field directional accuracy	50.7%	46.8%	-3.9pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	54.2%	49.2%	-4.9pp
Field sim win-rate	43.4%	40.0%	-3.4pp
Field sim net PnL	-\$337.72	-\$315.68	--

Week A is exactly the issue-#7 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, four windows back, inside issue #4's window). Week A (2026-09-14..2026-09-20) is pure grok-4.6 (it reproduces the published issue #7); week B is pure grok-4.6.

Practical implications

- Eight issues, eight margin-curve readings. The tier ordering moved again -- full unanimity took the top and the 4-peer tier fell from the top to the bottom -- and three of the four tiers printed a lower rate than in issue #7 (4 peers 55.5% -> 39.0%, full unanimity 51.6% -> 47.8%), so a consensus filter still should not carry weight in a config.
- The pooled lift is -14.0pp this week against +12.1pp in issue #7 and -30.4pp in issue #6. The sign flipped back below zero after one issue 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 -3.9pp week over week; 1 of the 4 consensus tiers moved up and 3 down against issue #7 (3 peers -13.0pp, 4 peers -16.5pp, 5 peers +2.6pp, unanimity -3.8pp).

Limitations

- Single week (Sep 21-27, 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 30 disagreeing calls; per-model counts run 0-16. 713 calls (14.4% of the 4,958-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window ran 2 trend days of 7 (issue #7: 3 of 7); every comparison with issue #7 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 8): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 21-27) daily coverage is FULL, as it was in issues #4, #5, #6 and #7: 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 488, 1M 485). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, four 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-#7 window), so every grok number in this issue and in issue #7 is a pure grok-4.6 line -- the grok week-over-week row is the third 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 #7 after the audit that found the raw table mixes two exchanges; issue #7 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 28 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6 and #7, which printed 52,971 at the Sep 21 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 57,929 is an additive step under one definition; counter deltas against issue #6 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,273
OK in gate	9,273	-- of them directional	4,958
Out of gate (1w / 1M)	488 / 485	-- of them sideways	4,315
Invalid	44	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-21.json	Report cutoff	Mon Sep 28, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-29 08:16 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,245 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 28, 2026 cutoff): 57,929 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 35 published reports.

Issue #8. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #7 asked "the pooled agree-vs-disagree lift flipped to +12.1pp after three issues below zero -- does the positive sign hold for a second issue?" -- it reads -14.0pp this week after +12.1pp in issue #7 (agree 46.0% vs disagree 60.0% on n=30), back below zero; and "The 4-peer tier has sat above the field base two issues running -- does it hold for a third?" -- 2 of the 4 tiers kept their issue-#7 side of the base, and the 4-peer tier came in below it. The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #7; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits four 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 went back below zero at -14.0pp after one issue above it -- does the negative sign hold for a second issue?
- The 4-peer tier fell below the field base after two issues above it -- does it stay below for a second issue?
- Monthly series: the agreement curve across regimes at monthly n.

Related research

Report	Direct PDF link
Weekly Calibration #8	https://marketmania.ai/research/reports/weekly-calibration-2026-09-21.pdf
Weekly Model Watch #8	https://marketmania.ai/research/reports/model-watch-2026-09-21.pdf
Consensus Watch #7	https://marketmania.ai/research/reports/consensus-watch-2026-09-14.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_2026w39,
  title = {Consensus Watch #8: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {30},
  url = {https://marketmania.ai/research/reports/consensus-watch-2026-09-21.pdf},
  note = {Methodology v1.1; window Sep 21-27, 2026 UTC; source weekly_metrics_2026-09-21.json}
}
```