

Consensus Watch #9

October 7, 2026 · MarketMania Research · Weekly series · window Sep 28-Oct 4, 2026 UTC

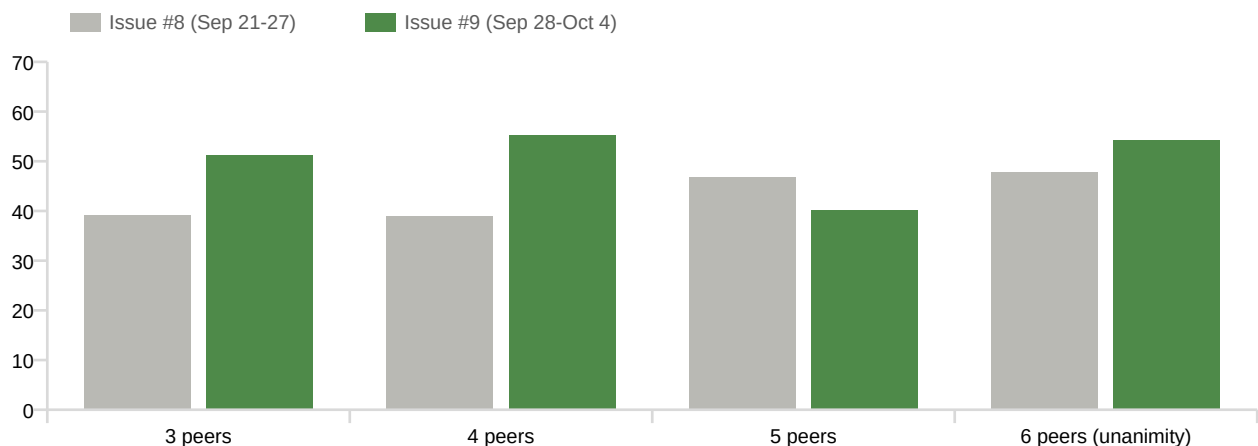
RESEARCH SNAPSHOT			
3,700 LOO-scoreable calls	7 models (stable lineup)	631 thin/tied excluded	window Sep 28-Oct 4, 2026 UTC
method = leave-one-out majority, 3+ directional peers	FH gate 1h / 4h / 1d	base field hit 49.6%	cutoff Mon Oct 5, 2026, 16:00 UTC
market: BTC net +2.44% (prior +4.06%)	ann vol 21.1% (was 52.2%)	TOP5 volume \$17.96B , -20% w/w	pairwise corr 0.57 (was 0.71)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 99.5% of leave-one-out scoreable calls and the pooled lift came in at -4.6pp: agree-hit 51.0% against 55.6% on disagreement, on a disagree bucket of n=18.

TL;DR

- **OBSERVATION -- 99.5% of the field's leave-one-out calls ran with the herd.** Of 3,700 LOO-scoreable directional calls, 3,682 (99.5%) sided with peers' majority (issue #8: 99.3%); agree-hit was 51.0% [49.4%, 52.6%] vs. 55.6% [33.7%, 75.4%] for the 18 disagreeers, a pooled lift of -4.6pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a ninth reading.** 3 peers 51.2% (was 39.1%), 4 peers 55.3% (was 39.0%), 5 peers 40.1% (was 46.8%), full unanimity 54.3% (was 47.8%). 3 of the 4 tiers sit above the 49.6% field base, and 2 of the 4 kept the side of the base they took in issue #8.
- **Daily-horizon unanimity split around the field base.** 1d/1d margin-6 hit 46.4% (n=28) and 1d/4h margin-6 hit 92.9% (n=14); issue #8 printed 51.7% and 39.6% on the same two cells.
- **Herd failures did not disappear.** 429 cells had 6+ directional models on one side (issue #8: 505); the 5 worst sat at 68.7-71.3 mean confidence and resolved 0.0%-33.3%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #9 n: 420 / 340 / 744 / 2,163; field base this week 49.6% (issue #8: 46.8%). A 2-peer bucket exists this week too (n=15, 20.0%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #8.

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 five 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 fourth 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	3,682	51.0%	[49.4%, 52.6%]	-4.6 (vs. disagree)
Disagree with LOO peers	18	55.6%	[33.7%, 75.4%]	ref.

Read with context. The disagree bucket is n=18 this week against 30 in issue #8 and 52 in issue #7; 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	2,535	50.8%	16	56.2%	-5.4	393
4h / 4h	447	52.8%	1	0.0%	+52.8	100
4h / 1h	614	49.8%	1	100.0%	-50.2	83
1d / 1d	54	48.1%	0	n/a	n/a	42
1d / 4h	32	65.6%	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 51.2% (n=420), 4 peers 55.3% (n=340), 5 peers 40.1% (n=744), 6 peers 54.3% (n=2,163). Against the 49.6% field base, 3 tier(s) sit above and 1 below, and 2 of the 4 tiers kept the side of the base they took in issue #8. Nine issues in, the margin curve has taken nine readings.

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

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 46.4% (n=28) and 1d/4h unanimity 92.9% (n=14), split around the 49.6% field base; the whole 1d/1d agree bucket ran 48.1% (n=54) -- not the highest of the five FH/TF cells -- while 1d/4h ran 65.6% (n=32), the thinnest agree bucket in the table above. Issue #8 printed 51.7% and 39.6% on the same two unanimity cells. Both daily buckets are more than an order of magnitude thinner than the hourly one (n=2,535), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 429 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #8: 505).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Oct 2, 12:01	BTC	4h / 4h	long	7	71.3	0.0%
Oct 4, 16:01	XRP	1h / 1h	long	7	69.3	14.3%
Oct 2, 19:01	BTC	1h / 1h	short	7	69.0	0.0%
Oct 4, 20:01	XRP	1h / 1h	long	7	68.9	28.6%
Oct 4, 08:01	XRP	4h / 1h	long	6	68.7	33.3%

The 5 worst cells were long/short swarms at 68.7-71.3 mean confidence and resolved 0.0%-33.3%. 'Sideways' was 53.4% of mature forecasts (4,955 of 9,286) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 0 to 8 this week (issue #8: 0 to 16). No line clears the N=10 floor, so every per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	552	51.3%	1	100.0%
claude-opus-5	503	51.3%	0	n/a
deepseek-v4-pro	563	50.6%	2	0.0%
gemini-3.1-pro	532	51.1%	8	62.5%
gpt-5.6-sol	553	49.7%	1	0.0%
grok-4.6	412	52.9%	3	100.0%
qwen-3.8-max	567	50.6%	3	33.3%

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 five 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 fourth pure-4.6 against pure-4.6 comparison of the series.

Market check: an up week on lower volatility

OBSERVATION — mean |1d move| 1.83% -> 0.87% (-52% rel), BTC realized vol 35.8% -> 32.2% (ann., hourly); field directional accuracy 46.8% -> 49.6% (+2.8pp), 5 of 7 models improved, sim win-rate up for 5 of 7, field sim PnL -\$315.68 -> -\$408.24 (adjacent calendar weeks Sep 21-27 vs Sep 28-Oct 4; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 49.2% -> 49.0% — the opposite direction to the trade-based hit rule.

Measure	Sep 21-27 (week A)	Sep 28-Oct 4 (week B)	Change
Mean 1d move (5 assets)	1.83%	0.87%	-52% rel
BTC realized vol (ann., hourly)	35.8%	32.2%	-3.6pp
Field directional accuracy	46.8%	49.6%	+2.8pp
Models improving hit-rate	--	5 of 7	--
Raw price-sign accuracy	49.2%	49.0%	-0.2pp
Field sim win-rate	40.0%	41.0%	+1.1pp
Field sim net PnL	-\$315.68	-\$408.24	--

Week A is exactly the issue-#8 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, five windows back, inside issue #4's window). Week A (2026-09-21..2026-09-27) is pure grok-4.6 (it reproduces the published issue #8); week B is pure grok-4.6.

Practical implications

- Nine issues, nine margin-curve readings. The tier ordering moved again -- the 4-peer tier went from the bottom to the top and the 5-peer tier fell to the bottom -- and three of the four tiers printed a higher rate than in issue #8 (4 peers 39.0% -> 55.3%, full unanimity 47.8% -> 54.3%), so a consensus filter still should not carry weight in a config.
- The pooled lift is -4.6pp this week against -14.0pp in issue #8 and +12.1pp in issue #7. 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 +2.8pp week over week; 3 of the 4 consensus tiers moved up and 1 down against issue #8 (3 peers +12.0pp, 4 peers +16.3pp, 5 peers -6.8pp, unanimity +6.5pp).

Limitations

- Single week (Sep 28-Oct 4, 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 18 disagreeing calls; per-model counts run 0-8. 631 calls (14.6% of the 4,331-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window ran 2 trend days of 7 (issue #8: 2 of 7); every comparison with issue #8 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 9): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 28-Oct 4) daily coverage is FULL, as it was in issues #4, #5, #6, #7 and #8: 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 486). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, five 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-#8 window), so every grok number in this issue and in issue #8 is a pure grok-4.6 line -- the grok week-over-week row is the fourth 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 #8 after the audit that found the raw table mixes two exchanges; issue #8 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 (Oct 5 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7 and #8, which printed 57,929 at the Sep 28 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 62,260 is an additive step under one definition; counter deltas against issue #7 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,286
OK in gate	9,286	-- of them directional	4,331
Out of gate (1w / 1M)	490 / 486	-- of them sideways	4,955
Invalid	28	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-28.json	Report cutoff	Mon Oct 5, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-07 06:16 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 3,700 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 5, 2026 cutoff): 62,260 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 43 published reports.

Issue #9. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #8 asked "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?" -- it reads -4.6pp this week after -14.0pp in issue #8 (agree 51.0% vs disagree 55.6% on n=18), below zero for a second issue; and "The 4-peer tier fell below the field base after two issues above it -- does it stay below for a second issue?" -- 2 of the 4 tiers kept their issue-#8 side of the base, and the 4-peer tier came in above it. The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #8; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits five 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 -4.6pp for a second issue -- does the negative sign hold for a third issue?
- The 4-peer tier went back above the field base after one issue 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 #9	https://marketmania.ai/research/reports/weekly-calibration-2026-09-28.pdf
Weekly Model Watch #9	https://marketmania.ai/research/reports/model-watch-2026-09-28.pdf
Consensus Watch #8	https://marketmania.ai/research/reports/consensus-watch-2026-09-21.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_2026w40,
  title = {Consensus Watch #9: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year = {2026}, month = {October}, day = {7},
  url = {https://marketmania.ai/research/reports/consensus-watch-2026-09-28.pdf},
  note = {Methodology v1.1; window Sep 28-Oct 4, 2026 UTC; source weekly_metrics_2026-09-28.json}
}
```