

Consensus Watch #10

October 14, 2026 · MarketMania Research · Weekly series · window Oct 5-11, 2026 UTC

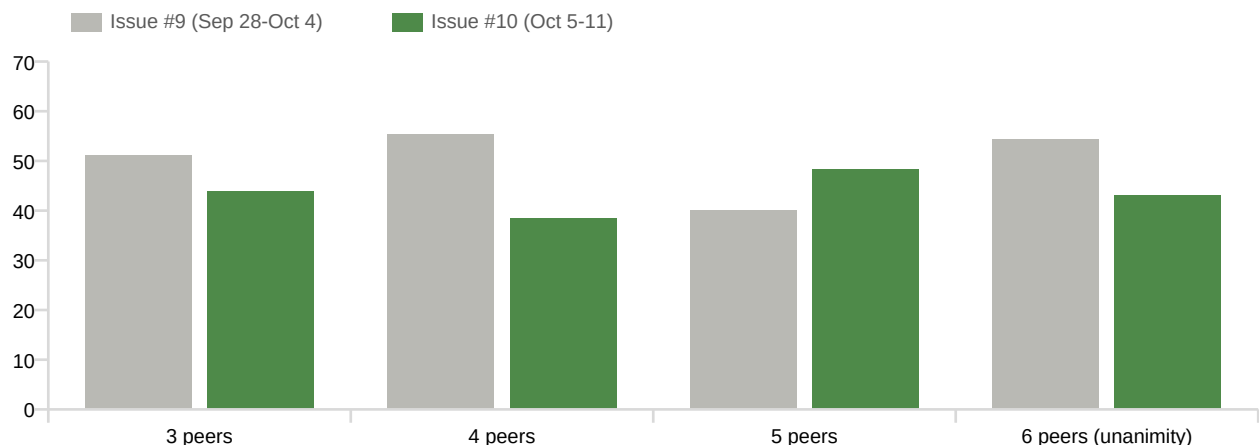
RESEARCH SNAPSHOT			
3,735 LOO-scoreable calls	7 models (stable lineup)	866 thin/tied excluded	window Oct 5-11, 2026 UTC
method = leave-one-out majority, 3+ directional peers	FH gate 1h / 4h / 1d	base field hit 43.8%	cutoff Mon Oct 12, 2026, 16:00 UTC
market: BTC net -3.19% (prior +2.44%)	ann vol 27.2% (was 21.1%)	TOP5 volume \$15.97B , -11% w/w	pairwise corr 0.85 (was 0.57)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 98.1% of leave-one-out scoreable calls and the pooled lift came in at +2.2pp: agree-hit 43.6% against 41.4% on disagreement, on a disagree bucket of n=70.

TL;DR

- **OBSERVATION -- 98.1% of the field's leave-one-out calls ran with the herd.** Of 3,735 LOO-scoreable directional calls, 3,665 (98.1%) sided with peers' majority (issue #9: 99.5%); agree-hit was 43.6% [42.0%, 45.2%] vs. 41.4% [30.6%, 53.1%] for the 70 disagreeers, a pooled lift of +2.2pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a tenth reading.** 3 peers 44.0% (was 51.2%), 4 peers 38.5% (was 55.3%), 5 peers 48.4% (was 40.1%), full unanimity 43.1% (was 54.3%). 2 of the 4 tiers sit above the 43.8% field base, and 1 of the 4 kept the side of the base they took in issue #9.
- **Daily-horizon unanimity fell below the field base.** 1d/1d margin-6 hit 25.7% (n=105) and 1d/4h margin-6 hit 16.7% (n=42); issue #9 printed 46.4% and 92.9% on the same two cells.
- **Herd failures did not disappear.** 372 cells had 6+ directional models on one side (issue #9: 429); the 5 worst sat at 68.1-70.9 mean confidence and resolved 0.0%-0.0%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #10 n: 452 / 605 / 858 / 1,708; field base this week 43.8% (issue #9: 49.6%). A 2-peer bucket exists this week too (n=42, 38.1%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #9.

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 six 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 fifth 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	3,665	43.6%	[42.0%, 45.2%]	+2.2 (vs. disagree)
Disagree with LOO peers	70	41.4%	[30.6%, 53.1%]	ref.

Read with context. The disagree bucket is n=70 this week against 18 in issue #9 and 30 in issue #8; 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	2,147	46.2%	43	44.2%	+2.1	549
4h / 4h	644	46.6%	16	37.5%	+9.1	137
4h / 1h	644	38.7%	9	33.3%	+5.3	125
1d / 1d	136	24.3%	0	n/a	n/a	24
1d / 4h	94	25.5%	2	50.0%	-24.5	31

"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 44.0% (n=452), 4 peers 38.5% (n=605), 5 peers 48.4% (n=858), 6 peers 43.1% (n=1,708). Against the 43.8% field base, 2 tier(s) sit above and 2 below, and 1 of the 4 tiers kept the side of the base they took in issue #9. Ten issues in, the margin curve has taken ten readings.

Daily-horizon herds: unanimity below the field base

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 25.7% (n=105) and 1d/4h unanimity 16.7% (n=42), both below the 43.8% field base; the whole 1d/1d agree bucket ran 24.3% (n=136) -- not the highest of the five FH/TF cells -- while 1d/4h ran 25.5% (n=94), the thinnest agree bucket in the table above. Issue #9 printed 46.4% and 92.9% on the same two unanimity cells. Both daily buckets are more than an order of magnitude thinner than the hourly one (n=2,147), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 372 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #9: 429).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Oct 5, 00:01	BTC	1d / 4h	long	7	70.9	0.0%
Oct 5, 00:01	BTC	4h / 4h	long	7	70.4	0.0%
Oct 5, 00:01	BNB	1d / 4h	long	7	69.0	0.0%
Oct 5, 00:01	BTC	1d / 1d	long	7	68.9	0.0%
Oct 5, 00:01	ETH	4h / 4h	long	7	68.1	0.0%

The 5 worst cells were long swarms at 68.1-70.9 mean confidence and resolved 0.0%-0.0%. 'Sideways' was 50.3% of mature forecasts (4,648 of 9,249) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 1 to 42 this week (issue #9: 0 to 8). 2 of the 7 lines clear the N=10 floor (deepseek-v4-pro 11, gemini-3.1-pro 42); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	517	44.3%	2	0.0%
claude-opus-5	474	46.2%	1	100.0%
deepseek-v4-pro	553	41.9%	11	54.5%
gemini-3.1-pro	491	42.2%	42	38.1%
gpt-5.6-sol	586	44.2%	5	20.0%
grok-4.6	493	43.8%	3	33.3%
qwen-3.8-max	551	43.0%	6	66.7%

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 six 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 fifth pure-4.6 against pure-4.6 comparison of the series.

Market check: a down week with bigger daily moves

OBSERVATION — mean |1d move| 0.87% -> 1.54% (+77% rel), BTC realized vol 32.2% -> 30.9% (ann., hourly); field directional accuracy 49.6% -> 43.8% (-5.9pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -\$408.24 -> -\$520.82 (adjacent calendar weeks Sep 28-Oct 4 vs Oct 5-11; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 49.0% -> 45.5% — the same direction as the trade-based hit rule.

Measure	Sep 28-Oct 4 (week A)	Oct 5-11 (week B)	Change
Mean 1d move (5 assets)	0.87%	1.54%	+77% rel
BTC realized vol (ann., hourly)	32.2%	30.9%	-1.3pp
Field directional accuracy	49.6%	43.8%	-5.9pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	49.0%	45.5%	-3.5pp
Field sim win-rate	41.0%	34.5%	-6.5pp
Field sim net PnL	-\$408.24	-\$520.82	--

Week A is exactly the issue-#9 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, six windows back, inside issue #4's window). Week A (2026-09-28..2026-10-04) is pure grok-4.6 (it reproduces the published issue #9); week B is pure grok-4.6.

Practical implications

- Ten issues, ten margin-curve readings. The tier ordering moved again -- the 5-peer tier went from the bottom to 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 #9 (4 peers 55.3% -> 38.5%, full unanimity 54.3% -> 43.1%), so a consensus filter still should not carry weight in a config.
- The pooled lift is +2.2pp this week against -4.6pp in issue #9 and -14.0pp in issue #8. The sign flipped back above zero after two issues below 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 -5.9pp week over week; 1 of the 4 consensus tiers moved up and 3 down against issue #9 (3 peers -7.2pp, 4 peers -16.8pp, 5 peers +8.3pp, unanimity -11.2pp).

Limitations

- Single week (Oct 5-11, 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 70 disagreeing calls; per-model counts run 1-42. 866 calls (18.8% of the 4,601-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window ran 2 trend days of 7 (issue #9: 2 of 7); every comparison with issue #9 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 10): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 5-11) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8 and #9: 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 479). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, six 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-#9 window), so every grok number in this issue and in issue #9 is a pure grok-4.6 line -- the grok week-over-week row is the fifth 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 #9 after the audit that found the raw table mixes two exchanges; issue #9 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 12 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8 and #9, which printed 62,260 at the Oct 5 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 66,861 is an additive step under one definition; counter deltas against issue #8 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,249
OK in gate	9,249	-- of them directional	4,601
Out of gate (1w / 1M)	490 / 479	-- of them sideways	4,648
Invalid	72	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-05.json	Report cutoff	Mon Oct 12, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-13 15:15 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 3,735 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 12, 2026 cutoff): 66,861 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 51 published reports.

Issue #10. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #9 asked "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?" -- it reads +2.2pp this week after -4.6pp in issue #9 (agree 43.6% vs disagree 41.4% on n=70), back above zero after two issues below it; and "The 4-peer tier went back above the field base after one issue below it -- does it stay above for a second issue?" -- 1 of the 4 tiers kept their issue-#9 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 #9; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits six 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 above zero at +2.2pp after two issues below it -- does the positive sign hold for a second issue?
- The 4-peer tier fell back below the field base after one issue 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 #10	https://marketmania.ai/research/reports/weekly-calibration-2026-10-05.pdf
Weekly Model Watch #10	https://marketmania.ai/research/reports/model-watch-2026-10-05.pdf
Consensus Watch #9	https://marketmania.ai/research/reports/consensus-watch-2026-09-28.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_2026w41,
  title = {Consensus Watch #10: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year = {2026}, month = {October}, day = {14},
  url = {https://marketmania.ai/research/reports/consensus-watch-2026-10-05.pdf},
  note = {Methodology v1.1; window Oct 5-11, 2026 UTC; source weekly_metrics_2026-10-05.json}
}
```