Consensus Watch #11

October 21, 2026 · MarketMania Research · Weekly series · window Oct 12-18, 2026 UTC

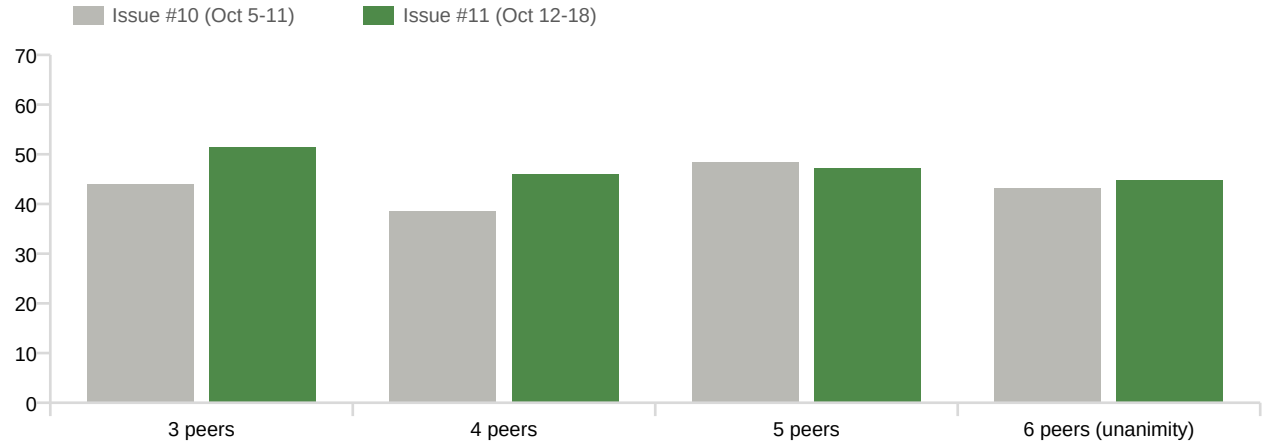
RESEARCH SNAPSHOT			
4,330 LOO-scoreable calls	**7** models (3 lines spliced)	**754** thin/tied excluded	window **Oct 12-18, 2026 UTC**
method = **leave-one-out** majority, 3+ directional peers	FH gate **1h / 4h / 1d**	base field hit **46.6%**	cutoff **Mon Oct 19, 2026, 16:00 UTC**
market: BTC net **-1.98%** (prior -3.19%)	ann vol **27.4%** (was 27.2%)	TOP5 volume **$15.84B**, -1% w/w	pairwise corr **0.94** (was 0.85)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 98.8% of leave-one-out scoreable calls and the pooled lift came in at +1.8pp: agree-hit 46.3% against 44.4% on disagreement, on a disagree bucket of n=54.

TL;DR

- **OBSERVATION -- 98.8% of the field's leave-one-out calls ran with the herd.** Of 4,330 LOO-scoreable directional calls, 4,276 (98.8%) sided with peers' majority (issue #10: 98.1%); agree-hit was 46.3% [44.8%, 47.8%] vs. 44.4% [32.0%, 57.6%] for the 54 disagreers, a pooled lift of +1.8pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took an eleventh reading.** 3 peers 51.4% (was 44.0%), 4 peers 46.0% (was 38.5%), 5 peers 47.1% (was 48.4%), full unanimity 44.8% (was 43.1%). 2 of the 4 tiers sit above the 46.6% field base, and 4 of the 4 kept the side of the base they took in issue #10.
- **Daily-horizon unanimity rose above the field base.** 1d/1d margin-6 hit 51.6% (n=91) and 1d/4h margin-6 hit 50.0% (n=84); issue #10 printed 25.7% and 16.7% on the same two cells.
- **Herd failures did not disappear.** 477 cells had 6+ directional models on one side (issue #10: 372); the 5 worst sat at 66.9-67.9 mean confidence and resolved 0.0%-14.3%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #11 n: 432 / 615 / 1,068 / 2,128; field base this week 46.6% (issue #10: 43.8%). A 2-peer bucket exists this week too (n=33, 54.5%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #10.

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. Each spliced line counts as one line, so no line counts as two peers in a cell. Three model lines flipped at Oct 17, 2026 09:01 UTC, INSIDE this window: gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5 and grok-4.6 -> grok-4.7. Every row of those three lines in this issue is a lineage splice (old-era rows Oct 12 00:00-Oct 17 09:01, new-era rows after), week A (the issue-#10 window) is entirely old-era, so each of the three week-over-week rows joins a spliced line to a single-build line. Engine ids are kept; the vendors' names are Claude Opus 5.5, GPT-6.1 Sol and Grok 4.7. claude-opus-5.5 (spliced) = claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC); gpt-6.1-sol (spliced) = gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC); grok-4.7 (spliced) = grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC).

Pooled result: agree vs. disagree with the LOO consensus

Group	n	Hit rate	Wilson 95% CI	Lift, pp
Agree with LOO peers	4,276	46.3%	[44.8%, 47.8%]	+1.8 (vs. disagree)
Disagree with LOO peers	54	44.4%	[32.0%, 57.6%]	ref.

Read with context. The disagree bucket is n=54 this week against 70 in issue #10 and 18 in issue #9; 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	2,559	48.5%	37	46.0%	+2.6	465
4h / 4h	718	41.5%	7	57.1%	-15.6	148
4h / 1h	700	40.6%	6	33.3%	+7.2	103
1d / 1d	142	50.0%	0	n/a	n/a	18
1d / 4h	157	53.5%	4	25.0%	+28.5	20

"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 51.4% (n=432), 4 peers 46.0% (n=615), 5 peers 47.1% (n=1,068), 6 peers 44.8% (n=2,128). Against the 46.6% 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 #10. Eleven issues in, the margin curve has taken eleven readings.

Daily-horizon herds: unanimity above the field base

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 51.6% (n=91) and 1d/4h unanimity 50.0% (n=84), both above the 46.6% field base; the whole 1d/1d agree bucket ran 50.0% (n=142) -- not the highest of the five FH/TF cells -- while 1d/4h ran 53.5% (n=157), the second-thinnest agree bucket in the table above. Issue #10 printed 25.7% and 16.7% on the same two unanimity cells. Both daily buckets are more than an order of magnitude thinner than the hourly one (n=2,559), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 477 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #10: 372).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Oct 15, 12:01	SOL	4h / 4h	short	7	67.9	0.0%
Oct 15, 04:01	BTC	4h / 1h	short	6	67.7	0.0%
Oct 14, 16:01	BNB	4h / 4h	short	7	67.6	14.3%
Oct 15, 04:01	XRP	4h / 4h	short	6	67.2	0.0%
Oct 15, 06:01	BTC	1h / 1h	short	7	66.9	0.0%

The 5 worst cells were short swarms at 66.9-67.9 mean confidence and resolved 0.0%-14.3%. 'Sideways' was 44.7% of mature forecasts (4,114 of 9,198) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 0 to 20 this week (issue #10: 1 to 42). 2 of the 7 lines clear the N=10 floor (deepseek-v4-pro 11, gemini-3.1-pro 20); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	605	46.1%	0	n/a
claude-opus-5.5 (spliced)	539	45.6%	5	100.0%
deepseek-v4-pro	623	47.8%	11	36.4%
gemini-3.1-pro	584	44.9%	20	45.0%
gpt-6.1-sol (spliced)	665	45.7%	9	22.2%
grok-4.7 (spliced)	615	46.5%	4	25.0%
qwen-3.8-max	645	47.1%	5	60.0%

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. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are lineage splices (the three lines flipped at Oct 17, 2026 09:01 UTC, inside this window; week A is entirely old-era) -- see How to read this.

Market check: bigger daily moves and higher BTC volatility

OBSERVATION — mean |1d move| 1.54% -> 2.05% (+33% rel), BTC realized vol 30.9% -> 34.9% (ann., hourly); field directional accuracy 43.8% -> 46.6% (+2.8pp), 6 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$520.82 -> -$352.09 (adjacent calendar weeks Oct 5-11 vs Oct 12-18; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 45.5% -> 44.1% — the opposite direction to the trade-based hit rule.

Measure	Oct 5-11 (week A)	Oct 12-18 (week B)	Change
Mean |1d move| (5 assets)	1.54%	2.05%	+33% rel
BTC realized vol (ann., hourly)	30.9%	34.9%	+3.9pp
Field directional accuracy	43.8%	46.6%	+2.8pp
Models improving hit-rate	--	6 of 7	--
Raw price-sign accuracy	45.5%	44.1%	-1.4pp
Field sim win-rate	34.5%	39.2%	+4.7pp
Field sim net PnL	-$520.82	-$352.09	--

Week A is exactly the issue-#10 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: the model cutover (2026-10-17T09:01:00+00:00, inside week B, day 6 of 7) splices 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 (2026-10-05..2026-10-11) is entirely before the cutover: its old-id rows are aggregated under the heads (it reproduces the published issue #10); week B mixes the old-id era with the new ids. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits seven windows back (inside issue #4's window).

Practical implications

- Eleven issues, eleven margin-curve readings. The tier ordering moved again -- the 3-peer tier went from second to the top and the 6-peer tier fell from third to the bottom -- and three of the four tiers printed a higher rate than in issue #10 (4 peers 38.5% -> 46.0%, full unanimity 43.1% -> 44.8%), so a consensus filter still should not carry weight in a config.
- The pooled lift is +1.8pp this week against +2.2pp in issue #10 and -4.6pp in issue #9. The sign stayed above 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 #10 (3 peers +7.4pp, 4 peers +7.5pp, 5 peers -1.3pp, unanimity +1.7pp).

Limitations

- Single week (Oct 12-18, 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 54 disagreeing calls; per-model counts run 0-20. 754 calls (14.8% of the 5,084-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window ran 3 trend days of 7 (issue #10: 2 of 7); every comparison with issue #10 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 11): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 12-18) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9 and #10: 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 483). (2) Model lineage is a splice. Three lines flipped at Oct 17, 2026 09:01 UTC, INSIDE this window (day 6 of 7): "claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the claude-opus-5 era, 315 from claude-opus-5.5); "gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the gpt-5.6-sol era, 315 from gpt-6.1-sol); "grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the grok-4.6 era, 315 from grok-4.7). Week A (the issue-#10 window) carries no new-build rows, so each of the three week-over-week rows joins a spliced line to a single-build line, and a remapped comparison is not the same model build. (3) The two hourly slots before the flip (Oct 17, 07:01 and 08:01 UTC) hold 60 invalid rows (claude-opus-5 20, gpt-5.6-sol 20, grok-4.6 20), counted in the invalid counter as they are; in those two hours the other 4 lines (claude-fable-5, deepseek-v4-pro, gemini-3.1-pro, qwen-3.8-max) carry ok rows. (4) deepseek-v4-pro carries 1,440 rows in this window against 1,470 for every other line (Oct 15: 180 rows instead of 210). (5) The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits seven 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 #10 after the audit that found the raw table mixes two exchanges; issue #10 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 19 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9 and #10, which printed 66,861 at the Oct 12 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 71,945 is an additive step under one definition; counter deltas against issue #9 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,260	Mature (scored pool)	9,198
OK in gate	9,198	-- of them directional	5,084
Out of gate (1w / 1M)	490 / 483	-- of them sideways	4,114
Invalid	89	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-12.json	Report cutoff	Mon Oct 19, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-21 03:36 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,330 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 19, 2026 cutoff): 71,945 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 55 published reports.

Issue #11. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #10 asked "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?" -- it reads +1.8pp this week after +2.2pp in issue #10 (agree 46.3% vs disagree 44.4% on n=54), above zero for a second issue; and "The 4-peer tier fell back below the field base after one issue above it -- does it stay below for a second issue?" -- 4 of the 4 tiers kept their issue-#10 side of the base, and the 4-peer tier came in below it. Three rows are lineage splices this issue (claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC); gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC); grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)) because the flip (Oct 17, 2026 09:01 UTC) fell inside the window; in issue #10 they were pure claude-opus-5, gpt-5.6-sol and grok-4.6. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits seven 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 above zero at +1.8pp for a second issue -- does the positive sign hold for a third issue?
- The 4-peer tier stayed below the field base for a second issue -- does it stay below for a third issue?
- Monthly series: the agreement curve across regimes at monthly n.

Related research

Report	Direct PDF link
Weekly Calibration #11	https://marketmania.ai/research/reports/weekly-calibration-2026-10-12.pdf
Weekly Model Watch #11	https://marketmania.ai/research/reports/model-watch-2026-10-12.pdf
Consensus Watch #10	https://marketmania.ai/research/reports/consensus-watch-2026-10-05.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_2026w42,
  title  = {Consensus Watch #11: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year   = {2026}, month = {October}, day = {21},
  url    = {https://marketmania.ai/research/reports/consensus-watch-2026-10-12.pdf},
  note   = {Methodology v1.1; window Oct 12-18, 2026 UTC; source weekly_metrics_2026-10-12.json}
}
```