Consensus Watch #12

October 28, 2026 · MarketMania Research · Weekly series · window Oct 19-25, 2026 UTC

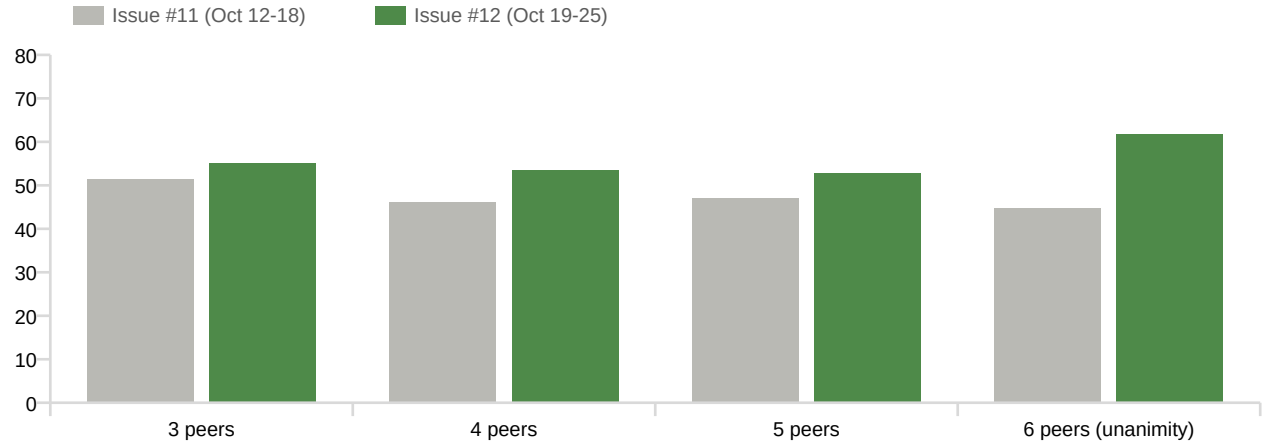
RESEARCH SNAPSHOT			
5,156 LOO-scoreable calls	**7** models (stable lineup)	**700** thin/tied excluded	window **Oct 19-25, 2026 UTC**
method = **leave-one-out** majority, 3+ directional peers	FH gate **1h / 4h / 1d**	base field hit **57.0%**	cutoff **Mon Oct 26, 2026, 16:00 UTC**
market: BTC net **-2.39%** (prior -1.98%)	ann vol **29.3%** (was 27.4%)	TOP5 volume **$14.02B**, -12% w/w	pairwise corr **0.93** (was 0.94)

KEY FINDING · OBSERVATION (one weekly window)

Herding ran at 98.9% of leave-one-out scoreable calls and the pooled lift came in at +7.0pp: agree-hit 57.9% against 50.8% on disagreement, on a disagree bucket of n=59.

TL;DR

- **OBSERVATION -- 98.9% of the field's leave-one-out calls ran with the herd.** Of 5,156 LOO-scoreable directional calls, 5,097 (98.9%) sided with peers' majority (issue #11: 98.8%); agree-hit was 57.9% [56.5%, 59.2%] vs. 50.8% [38.4%, 63.2%] for the 59 disagreers, a pooled lift of +7.0pp. Descriptive, one week, not a claim of a durable edge.
- **The margin curve took a twelfth reading.** 3 peers 55.0% (was 51.4%), 4 peers 53.5% (was 46.0%), 5 peers 52.8% (was 47.1%), full unanimity 61.7% (was 44.8%). 1 of the 4 tiers sit above the 57.0% field base, and 1 of the 4 kept the side of the base they took in issue #11.
- **Daily-horizon unanimity stayed above the field base.** 1d/1d margin-6 hit 100.0% (n=21) and 1d/4h margin-6 hit 94.5% (n=91); issue #11 printed 51.6% and 50.0% on the same two cells.
- **Herd failures did not disappear.** 588 cells had 6+ directional models on one side (issue #11: 477); the 5 worst sat at 67.0-70.1 mean confidence and resolved 0.0%-14.3%.



Hit rate (%) by number of agreeing LOO peers, week over week. Issue #12 n: 480 / 620 / 1,128 / 2,821; field base this week 57.0% (issue #11: 46.6%). A 2-peer bucket exists this week too (n=48, 39.6%) -- footnote only, far too small to plot. "Was" values are the numbers published in issue #11.

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. Week A (Oct 12-18, the issue-#11 window) is the spliced week: the three lines flipped at Oct 17, 2026 09:01 UTC inside it (gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5, grok-4.6 -> grok-4.7), so each week-over-week row for those lines joins this week's single-build line to last week's spliced line. This window (Oct 19-25) is the first full week on the new builds. In issue #11 those lines carried the labels "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)" and "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	5,097	57.9%	[56.5%, 59.2%]	+7.0 (vs. disagree)
Disagree with LOO peers	59	50.8%	[38.4%, 63.2%]	ref.

Read with context. The disagree bucket is n=59 this week against 54 in issue #11 and 70 in issue #10; 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	3,246	55.5%	45	53.3%	+2.1	424
4h / 4h	807	56.5%	5	40.0%	+16.5	111
4h / 1h	797	64.5%	4	75.0%	-10.5	126
1d / 1d	96	72.9%	1	0.0%	+72.9	23
1d / 4h	151	72.9%	4	25.0%	+47.8	16

"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 55.0% (n=480), 4 peers 53.5% (n=620), 5 peers 52.8% (n=1,128), 6 peers 61.7% (n=2,821). Against the 57.0% field base, 1 tier(s) sit above and 3 below, and 1 of the 4 tiers kept the side of the base they took in issue #11. Twelve issues in, the margin curve has taken twelve readings.

Daily-horizon herds: unanimity above the field base

The two daily cells stay the report's thinnest. 1d/1d unanimity hit 100.0% (n=21) and 1d/4h unanimity 94.5% (n=91), both above the 57.0% field base; the whole 1d/1d agree bucket ran 72.9% (n=96) -- the highest of the five FH/TF cells -- while 1d/4h ran 72.9% (n=151), the second-thinnest agree bucket in the table above. Issue #11 printed 51.6% and 50.0% on the same two unanimity cells. Both daily buckets are more than an order of magnitude thinner than the hourly one (n=3,246), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This week 588 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side (issue #11: 477).

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Oct 21, 16:01	SOL	4h / 1h	long	7	70.1	0.0%
Oct 21, 16:01	SOL	1h / 1h	long	7	68.4	0.0%
Oct 21, 16:01	BTC	4h / 4h	long	7	68.1	0.0%
Oct 21, 18:01	SOL	1h / 1h	long	7	67.1	14.3%
Oct 21, 16:01	ETH	4h / 4h	long	7	67.0	0.0%

The 5 worst cells were long swarms at 67.0-70.1 mean confidence and resolved 0.0%-14.3%. 'Sideways' was 36.1% of mature forecasts (3,315 of 9,171) this week.

Per-model: agree vs. disagree

Per-model disagree counts run 4 to 16 this week (issue #11: 0 to 20). 4 of the 7 lines clear the N=10 floor (deepseek-v4-pro 11, gemini-3.1-pro 16, gpt-6.1-sol 10, grok-4.7 10); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	698	57.3%	4	25.0%
claude-opus-5.5	651	57.9%	4	100.0%
deepseek-v4-pro	759	57.6%	11	54.5%
gemini-3.1-pro	734	60.2%	16	62.5%
gpt-6.1-sol	797	58.6%	10	30.0%
grok-4.7	712	57.3%	10	40.0%
qwen-3.8-max	746	56.2%	4	50.0%

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 week-over-week rows join this week's single-build line to issue #11's spliced line (the three lines flipped at Oct 17, 2026 09:01 UTC, inside week A) -- see How to read this.

Market check: smaller daily moves on lower volume

OBSERVATION — mean |1d move| 2.05% -> 1.35% (-34% rel), BTC realized vol 34.9% -> 27.5% (ann., hourly); field directional accuracy 46.6% -> 57.0% (+10.4pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$352.09 -> -$43.00 (adjacent calendar weeks Oct 12-18 vs Oct 19-25; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 44.1% -> 56.0% — the same direction as the trade-based hit rule.

Measure	Oct 12-18 (week A)	Oct 19-25 (week B)	Change
Mean |1d move| (5 assets)	2.05%	1.35%	-34% rel
BTC realized vol (ann., hourly)	34.9%	27.5%	-7.4pp
Field directional accuracy	46.6%	57.0%	+10.4pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	44.1%	56.0%	+11.9pp
Field sim win-rate	39.2%	46.6%	+7.5pp
Field sim net PnL	-$352.09	-$43.00	--

Week A is exactly the issue-#11 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 note: the model cutover (2026-10-17T09:01:00+00:00, inside week A, 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 (2026-10-12..2026-10-18, the published issue-#11 window) holds

the cutover: its old-id rows are aggregated under the heads (it reproduces the published issue #11); week B (2026-10-19..2026-10-25) is the first full week on the heads (0 old-id row(s) remapped in it). The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits eight windows back (inside issue #4's window).

Practical implications

- Twelve issues, twelve margin-curve readings. The tier ordering moved again -- the 6-peer tier went from the bottom to the top and the 5-peer tier fell from second to the bottom -- and all four tiers printed a higher rate than in issue #11 (4 peers 46.0% -> 53.5%, full unanimity 44.8% -> 61.7%), so a consensus filter still should not carry weight in a config.
- The pooled lift is +7.0pp this week against +1.8pp in issue #11 and +2.2pp in issue #10. The sign stayed above zero for a third 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 +10.4pp week over week; 4 of the 4 consensus tiers moved up and 0 down against issue #11 (3 peers +3.6pp, 4 peers +7.5pp, 5 peers +5.6pp, unanimity +16.9pp).

Limitations

- Single week (Oct 19-25, 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 59 disagreeing calls; per-model counts run 4-16. 700 calls (12.0% of the 5,856-call mature directional universe) were thin or tied and are excluded from every consensus table.
- This window had trend on 1 of 7 days (issue #11: 3 of 7); every comparison with issue #11 is a comparison across regimes as well as across weeks. Series density and lineage notes (wave 12): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 19-25) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10 and #11: 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 480, 1M 478). (2) This window is the first full week on the builds that took over at Oct 17, 2026 09:01 UTC: claude-opus-5.5, gpt-6.1-sol and grok-4.7 carry no rows from the old builds, and all 7 lines carry 1,470 rows each. Week A (the issue-#11 window) is the spliced week -- "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) -- so each of the three week-over-week rows joins this week's single-build line to last week's spliced line, and a remapped comparison is not the same model build. (3) No roll-out slot of the flip falls inside this window: the two hourly slots before it (Oct 17, 07:01 and 08:01 UTC) lie in week A, so this window's invalid counter holds none of their rows. (4) Past lineage flips: the Oct 17, 2026 09:01 UTC flip sits one window back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eight 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 #11 after the audit that found the raw table mixes two exchanges; issue #11 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 26 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10 and #11, which printed 71,945 at the Oct 19 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 77,801 is an additive step under one definition; counter deltas against issue #10 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,171
OK in gate	9,171	-- of them directional	5,856
Out of gate (1w / 1M)	480 / 478	-- of them sideways	3,315
Invalid	161	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-19.json	Report cutoff	Mon Oct 26, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-27 11:37 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,156 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 26, 2026 cutoff): 77,801 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 59 published reports.

Issue #12. Consensus Watch is a living weekly comparison; each issue appends one more week of leave-one-out agreement-vs-hit data. Issue #11 asked "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?" -- it reads +7.0pp this week after +1.8pp in issue #11 (agree 57.9% vs disagree 50.8% on n=59), above zero for a third issue; and "The 4-peer tier stayed below the field base for a second issue -- does it stay below for a third issue?" -- 1 of the 4 tiers kept their issue-#11 side of the base, and the 4-peer tier came in below it. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the first full week on the Oct 17 builds; week-over-week rows compare against the spliced issue-#11 week, where those lines carried the labels "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)" and "grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)". The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eight 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 +7.0pp for a third issue -- does the positive sign hold for a fourth issue?
- The 4-peer tier stayed below the field base for a third issue -- does it stay below for a fourth issue?
- Monthly series: the agreement curve across regimes at monthly n.

Related research

Report	Direct PDF link
Weekly Calibration #12	https://marketmania.ai/research/reports/weekly-calibration-2026-10-19.pdf
Weekly Model Watch #12	https://marketmania.ai/research/reports/model-watch-2026-10-19.pdf
Consensus Watch #11	https://marketmania.ai/research/reports/consensus-watch-2026-10-12.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_2026w43,
  title  = {Consensus Watch #12: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year   = {2026}, month = {October}, day = {28},
  url    = {https://marketmania.ai/research/reports/consensus-watch-2026-10-19.pdf},
  note   = {Methodology v1.1; window Oct 19-25, 2026 UTC; source weekly_metrics_2026-10-19.json}
}
```