

Monthly Consensus Watch #1

September 4, 2026 · MarketMania Research · Monthly series · window Aug 1-31, 2026 UTC

RESEARCH SNAPSHOT			
16,089 LOO-scoreable calls	7 models (lineage spliced: grok, qwen)	3,294 thin/tied excluded	window Aug 1-31, 2026 UTC
method = leave-one-out majority, 3+ directional peers	FH gate 1h / 4h / 1d	base field hit 46.3%	cutoff Mon Sep 1, 2026, 16:00 UTC
market: BTC net +24.95% (prior -3.31%)	ann vol 41.5% (was 27.3%)	TOP5 volume \$65.97B, +136% m/m	pairwise corr 0.79 (was 0.79)

KEY FINDING · OBSERVATION (one monthly window)

Herdling ran at 98.7% of leave-one-out scoreable calls across the month and the pooled lift came in at +1.9pp: agree-hit 45.5% against 43.6% on disagreement, on a disagree bucket of n=211.

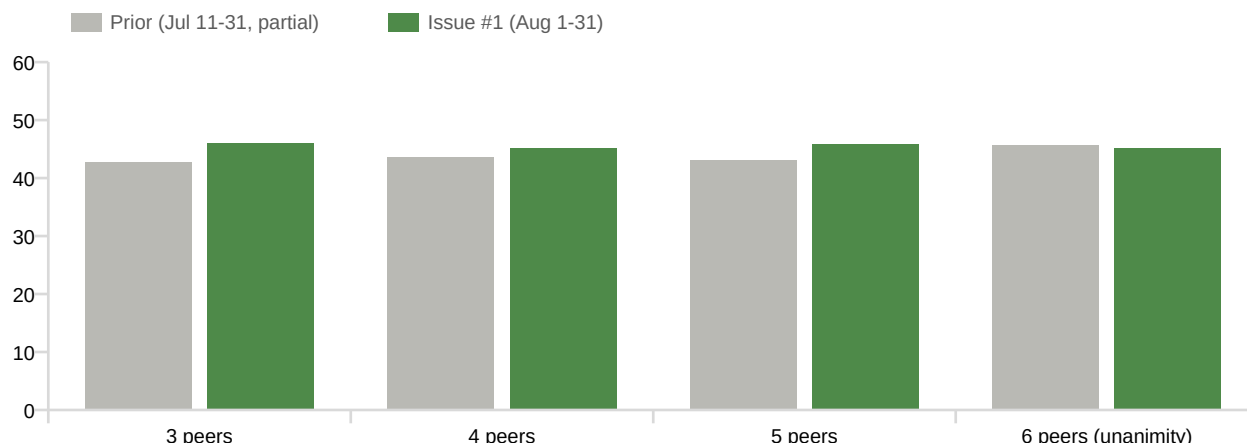
WEEKLY EVIDENCE -> MONTHLY VERDICT (rule below)

Metric	W1 Aug 3-9	W2 Aug 10-16	W3 Aug 17-23	W4 Aug 24-30	Month	Verdict
LOO lift, pp	-22.8	-6.9	+17.1	-6.0	+1.9	Rejected (1/4)
Herd share >= 95%	99.4%	99.8%	97.1%	99.1%	98.7%	Confirmed (4/4)
Unanimity hit vs field hit	-8.9	-0.5	+2.5	-3.5	-1.2	Confirmed (3/4)

Verdict rule: each row states one reading of the month; a week agrees when it shows the same reading (same sign, or the same name). Confirmed = 3 or 4 of the 4 weeks agree with the month; Inconclusive = 2; Rejected = 0 or 1. Weekly cells come from the frozen weekly packs (7-line roster, lineage spliced); the month is the pooled Aug 1-31 pull.

TL;DR

- **OBSERVATION -- 98.7% of the field's leave-one-out calls ran with the herd over the month.** Of 16,089 LOO-scoreable directional calls, 15,878 (98.7%) sided with peers' majority; agree-hit was 45.5% [44.8%, 46.3%] vs. 43.6% [37.1%, 50.3%] for the 211 disagreeers, a pooled lift of +1.9pp. Descriptive, one month, not a claim of a durable edge.
- **The margin curve read 45.9% / 45.1% / 45.8% / 45.1% from 3 peers to unanimity.** 0 of the 4 tiers sit above the 46.3% field base. The same four tiers read 42.7% / 43.5% / 43.1% / 45.6% in the prior window Jul 11-31 (partial), on n=1,609 / 2,394 / 3,628 / 11,531 there.
- **Week by week the lift changed sign 2 times.** The four published weeks ran Aug 3-9 -22.8pp, Aug 10-16 -6.9pp, Aug 17-23 +17.1pp, Aug 24-30 -6.0pp; the month pools them at +1.9pp and the prior window read -10.6pp. The month is a weighted pool of the four weeks, not their mean: the weekly range is -22.8pp to +17.1pp.
- **1,672 cells had 6+ directional models on one side across the month.** The 5 worst sat at 69.9-71.7 mean confidence and resolved 0.0%-0.0%; the prior window Jul 11-31 (partial) had 1,456 such cells, hitting 45.6% on n=11,531.



Hit rate (%) by number of agreeing LOO peers, the prior window against the whole month. Monthly n: 3 peers 1,964 / 4 peers 2,630 / 5 peers 4,482 / 6 peers 6,664. Prior n: 3 peers 1,609 / 4 peers 2,394 / 5 peers 3,628 / 6 peers 11,531. Field base this month 46.3%, prior window 45.0%. The prior window is PARTIAL: window A of the same pull, slots 2026-07-01 to 2026-08-01, forecasts counted from Jul 11 (first slot actually seen 2026-07-15 18:01 UTC) and audited candles from Jul 15, matured at the prior cutoff Sat Aug 1, 2026, 16:00 UTC under the same engine, the same FH gate and the same hit rule as August -- 19,541 directional calls against 19,383 in the month.

Why it matters

MarketMania publishes forecasts from 7 frontier model lines into the same slots every hour; 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. The four weekly issues inside this month answered it with lifts of Aug 3-9 -22.8pp, Aug 10-16 -6.9pp, Aug 17-23 +17.1pp, Aug 24-30 -6.0pp, and the pooled month reads +1.9pp. This monthly issue asks the same question at monthly n, across four weeks the pack labels as 2 calm and 2 volatile, using the same leave-one-out design built to avoid 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. 2 model lines are lineage splices this month (grok-4.6 (incl. 4.5-era, Aug 1-24); qwen-3.8-max (incl. 3.7-era, Aug 1-5)), and a spliced line never counts as two peers in a cell. The month is one pull at one cutoff (Mon Sep 1, 2026, 16:00 UTC); the four weekly blocks quoted alongside it are the published issues at their own frozen cutoffs.

Pooled result: agree vs. disagree with the LOO consensus

Group	n	Hit rate	Wilson 95% CI	Lift, pp
Agree with LOO peers	15,878	45.5%	[44.8%, 46.3%]	+1.9 (vs. disagree)
Disagree with LOO peers	211	43.6%	[37.1%, 50.3%]	ref.

The disagree bucket is n=211 over the whole month; the two Wilson intervals overlap. Prior window Jul 11-31 (partial): agree 44.6% / disagree 55.3%, lift -10.6pp, on n=16,557 agree and n=436 disagree. These are dependent observations pooled across four weeks and two volatility regimes.

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	9,857	45.2%	125	48.0%	-2.8	2,128
4h / 4h	2,483	46.1%	54	38.9%	+7.2	549
4h / 1h	2,548	44.9%	25	40.0%	+4.9	466
1d / 1d	454	54.0%	5	20.0%	+34.0	80
1d / 4h	536	45.3%	2	0.0%	+45.3	71

"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 month-over-month chart is on page 1; the full per-tier numbers: 3 peers 45.9% (n=1,964), 4 peers 45.1% (n=2,630), 5 peers 45.8% (n=4,482), 6 peers 45.1% (n=6,664). Against the 46.3% field base, 0 tiers sit above and 4 tiers below; in the prior window the same tiers read 42.7% / 43.5% / 43.1% / 45.6%. The week-by-week table further down prints each tier for each of the four weeks.

Daily-horizon herds: the thinnest cell, the best rate

The two daily cells stay the report's thinnest even at monthly n. 1d/1d unanimity hit 60.2% (n=161) and 1d/4h unanimity 46.0% (n=259) against the 46.3% field base; the whole 1d/1d agree bucket ran 54.0% (n=454) -- the highest of the five FH/TF cells -- while 1d/4h ran 45.3% (n=536). Both daily buckets are an order of magnitude thinner than the hourly one (n=9,857), so they are reported, not ranked.

Herd failures: the 5 worst unanimous misses

When the whole field agreed, and was wrong. This month 1,672 slot / symbol / FH / TF cells had 6 or more directional models sitting on the same side.

Slot (UTC)	Symbol	FH / TF	Side	Models	Mean conf	Hit rate
Aug 4, 19:01	BTC	1h / 1h	long	7	71.7	0.0%
Aug 11, 12:01	BNB	4h / 1h	long	7	71.0	0.0%
Aug 2, 06:01	SOL	1h / 1h	long	6	70.8	0.0%
Aug 3, 16:01	XRP	4h / 1h	long	7	70.0	0.0%
Aug 5, 20:01	BTC	4h / 1h	long	7	69.9	0.0%

The 5 worst cells were long swarms at 69.9-71.7 mean confidence and resolved 0.0%-0.0%. 'Sideways' was 52.7% of mature forecasts (21,596 of 40,979) this month.

Per-model: agree vs. disagree

Per-model disagree counts run 4 to 89 over the month. 5 of the 7 lines clear the N=10 floor (claude-fable-5 13, deepseek-v4-pro 39, gemini-3.1-pro 89, grok-4.6 (spliced) 43, qwen-3.8-max (spliced) 15); every other per-model contrarian cell below is insufficient.

Model	Agree n	Agree hit	Disagree n	Disagree hit
claude-fable-5	2,391	45.8%	13	53.8%
claude-opus-5	2,105	47.3%	4	50.0%
deepseek-v4-pro	2,065	44.9%	39	41.0%
gemini-3.1-pro	2,249	45.4%	89	47.2%
gpt-5.6-sol	2,459	44.5%	8	37.5%
grok-4.6 (spliced)	2,218	46.0%	43	44.2%
qwen-3.8-max (spliced)	2,391	45.0%	15	20.0%

Prior window Jul 11-31 (partial): per-model agree-side hit ran 44.1% to 45.2% on n=1,651 to 2,852 per line. At these counts a handful of flipped calls moves a per-model lift by tens of points, so models are not ranked by contrarian lift here. grok-4.6 (spliced) = grok-4.6 (incl. 4.5-era, Aug 1-24). qwen-3.8-max (spliced) = qwen-3.8-max (incl. 3.7-era, Aug 1-5).

Week by week: how often the field agreed, and what it paid

The same leave-one-out scoring run on each of the four Mon-Sun weeks of August gives four weekly lifts, and the pooled monthly number is their call-weighted mix. Herd share ran 97.1% to 99.8% and the lift ran -22.8pp to +17.1pp across the four weeks against +1.9pp for the month as a whole.

Week	Scoreable	Herd share	Agree hit	Disagree hit	Lift pp	Unanimity n	Unanimity hit
W1 Aug 3-9	3,327	99.4%	40.4%	63.2%	-22.8	2,380	41.5%
W2 Aug 10-16	2,929	99.8%	43.1%	50.0%	-6.9	1,961	42.4%
W3 Aug 17-23	4,357	97.1%	54.8%	37.6%	+17.1	2,887	55.4%
W4 Aug 24-30	3,884	99.1%	41.2%	47.2%	-6.0	2,709	38.7%
Month	16,089	98.7%	45.5%	43.6%	+1.9	10,984	45.4%

Source: the month pull split by slot week at the single monthly cutoff (Mon Sep 1, 2026, 16:00 UTC), so the four rows and the Month row are one consistent pool. The frozen-cutoff cross-check is the published weekly issues at their own cutoffs: the recomputed lift reproduces the published lift for 4 of 4 weeks (W1 -22.8pp vs -22.8pp, W2 -6.9pp vs -6.9pp, W3 +17.1pp vs +17.1pp, W4 -6.0pp vs -6.0pp). * = N<10 (insufficient) -- such cells are never ranked.

Margin	W1	W2	W3	W4	Month
3 peers	30.4% (404)	50.2% (428)	56.4% (468)	49.3% (456)	45.9% (1,964)
4 peers	43.3% (515)	39.4% (525)	48.9% (700)	47.6% (620)	45.1% (2,630)
5 peers	49.9% (1,140)	41.0% (792)	54.4% (1,128)	35.9% (924)	45.8% (4,482)
6 peers (unanimity)	33.6% (1,246)	43.3% (1,169)	56.5% (1,855)	39.4% (1,827)	45.1% (6,664)

Cells are hit rate with n in parentheses; * = N<10 (insufficient), "n/a (0)" = no call landed in that tier that week. Weeks are W1 Aug 3-9, W2 Aug 10-16, W3 Aug 17-23, W4 Aug 24-30. Same source and cutoff as the table above; the Month column is the pooled month, not the mean of the four weeks. Month-over-month columns compare back-to-back windows: Jul 11-31 (partial) vs Aug 1-31 (this issue). "Was" values are the pack's own prior-window block -- window A of the same pull, same engine, same FH gate and same hit rule, matured at the prior cutoff Sat Aug 1, 2026, 16:00 UTC. The prior month is PARTIAL: forecasts run from Jul 11 and audited candles from Jul 15, so it holds 19,541 directional calls against 19,383 in August. Deltas are computed on unrounded rates.

Market check: August ran hotter than the partial July

OBSERVATION — mean [1d move] 1.54% -> 1.98% (+29% rel), BTC realized vol 33.5% -> 36.9% (ann., hourly); field directional accuracy 45.0% -> 46.3% (+1.4pp), 5 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -\$3095.02 -> -\$1054.79 (Jul 11-31 (partial) vs Aug 1-31; descriptive, one pair of windows, not a claim). Robustness: raw price-sign accuracy 46.8% -> 46.0% — the same direction as the trade-based hit rule.

Measure	Jul 11-31 (partial)	Aug 1-31	Change
Mean [1d move] (5 assets)	1.54%	1.98%	+29% rel
BTC realized vol (ann., hourly)	33.5%	36.9%	+3.5pp
Field directional accuracy	45.0%	46.3%	+1.4pp
Models improving hit-rate	--	5 of 7	--
Raw price-sign accuracy	46.8%	46.0%	-0.9pp
Field sim win-rate	31.9%	35.8%	+3.9pp
Field sim net PnL	-\$3095.02	-\$1054.79	--

Window A is the pack's prior window -- a PARTIAL July (forecasts from Jul 11), not a full month -- so this slice and every month-over-month column in this report compare a full August against a partial July. Descriptive, one pair of windows; 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 splices in this window: grok-4.6 (incl. 4.5-era, Aug 1-24); qwen-3.8-max (incl. 3.7-era, Aug 1-5). 5,390 rows were remapped inside the month against 19,532 in the prior window.

Practical implications

- 0 of the 4 margin tiers sit above the 46.3% field base this month; the four tier rates span 45.1% to 45.9%, a spread of +0.8pp, and the four weeks inside the month ran lifts of -22.8pp to +17.1pp against a pooled monthly lift of +1.9pp.
- The pooled monthly lift is +1.9pp while the four weeks ran -22.8pp to +17.1pp and the prior window read -10.6pp.
- The field base moved +1.4pp from the partial July to August (45.0% -> 46.3%), and every consensus tier is measured against that base.

Limitations

- The prior window is a PARTIAL month: forecasts run only from Jul 11 and the audited daily candles only from Jul 15, so every "prior" column, the gray series of the page-1 chart and the market-check A column describe Jul 11-31 (partial), not a full July. It is window A of the same pull, slots 2026-07-01 to 2026-08-01, forecasts counted from Jul 11 (first slot actually seen 2026-07-15 18:01 UTC) and audited candles from Jul 15, matured at the prior cutoff Sat Aug 1, 2026, 16:00 UTC under the same engine, the same FH gate and the same hit rule as August -- 19,541 directional calls against 19,383 in the month. Month-over-month deltas are a step between two windows of different length.
- Two cutoffs live in this issue. The month is frozen at Mon Sep 1, 2026, 16:00 UTC; the four weekly blocks are frozen at their own published cutoffs (W1 Aug 3-9, W2 Aug 10-16, W3 Aug 17-23, W4 Aug 24-30). A forecast still pending at its week's cutoff is immature in the weekly pack and mature in the month pull, so the month does not equal the sum of the weeks (19,383 vs 17,418 directional, plus 1,965 on the 3 remainder days) -- that gap is arithmetic, not a data problem.
- Observations are not independent: the same models watch overlapping symbol / FH / TF cells hour after hour for 31 days. Wilson intervals here are descriptive, not inferential, and pooling a month does not make them inferential.
- The pooled lift rests on 211 disagreeing calls; per-model counts run 4-89. 3,294 calls (17.0% of the 19,383-call mature directional universe) were thin or tied and are excluded from every consensus table. Correlational, not causal: LOO removes self-match bias, not the confound that an easy slot lifts agreement and hit together.
- This window ran 9 trend days of 31, and the pack's own volatility rule splits the four Mon-Sun weeks as W1 calm, W2 calm, W3 volatile, W4 volatile; every month-over-month comparison is a comparison across regimes as well as across windows. Series density and lineage notes (monthly #1): (1) rolling-1w forecasts emit daily

since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so across this 31-day window daily coverage is PARTIAL as measured -- the 1w series has slots on 13 of 31 days and the 1M series on 10 of 31 days; both sit outside this report's FH gate (1h/4h/1d) and appear only in the exclusions counter (1w 906, 1M 698). (2) 2 model lines are lineage splices inside this window: grok-4.6 (incl. 4.5-era, Aug 1-24); qwen-3.8-max (incl. 3.7-era, Aug 1-5). (3) Every per-model aggregate in this issue sees the merged head id; the raw id is kept in the pack, and 5,390 rows were remapped inside the month against 19,532 in the prior window.

- Market-state row is computed on a single-exchange (binance) daily candle series, as in the weekly series after the audit that found the raw table mixes two exchanges. The prior-month block is PARTIAL: audited daily candles start 2026-07-15, so it covers 17 days (Jul 15-31), and the TOP5 volume figure compares a 31-day sum with a 17-day sum -- a level difference, not a like-for-like change.
- Research-to-date counter: directional forecasts resolved inside the FH gate (1h/4h/1d) since Jul 11, counted at the issue cutoff (Mon Sep 1, 2026, 16:00 UTC) -- the D2 definition pinned in the weekly series; this issue prints 39,073. The pack carries no prior-cutoff pin for a monthly window (the prior-cutoff pin is null by construction), so the control this issue is the weekly reproduction: the W4 block of this pack reproduces the published weekly #4 on 7 of 7 checks. Counter deltas against pre-wave-3 issues remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	42,654	Mature (scored pool)	40,979
OK in gate	40,979	-- of them directional	19,383
Out of gate (1w / 1M)	906 / 698	-- of them sideways	21,596
Invalid	71	Pending (next issue)	0
Late closes	0	Uptime, all grids	1,173 / 1,178 slots
Source file	monthly_metrics_2026-08.json	Report cutoff	Mon Sep 1, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-03 12:43 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 16,089 LOO-scored observations. MARKETMANIA RESEARCH TO DATE (as of cutoff Sep 1, 16:00 UTC): 39,073 resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 19 published reports.

Issue #1. Monthly Consensus Watch is a living monthly comparison; each issue appends one more calendar month of leave-one-out agreement-vs-hit data. This first issue sets the baseline: herd share 98.7%, pooled lift +1.9pp (agree 45.5% vs disagree 43.6% on n=211), four weekly lifts inside the month running -22.8pp to +17.1pp, and a prior window Jul 11-31 (partial) at herd share 97.4% and lift -10.6pp. 2 model lines are lineage splices this month (grok-4.6 (incl. 4.5-era, Aug 1-24); qwen-3.8-max (incl. 3.7-era, Aug 1-5)). Engine 1.1 has powered the sandbox since Aug 18, i.e. inside this window; no cross-engine PnL comparisons are claimed.

What we're testing next

- Monthly #2 (September): the same three series over Sep 1-30, cut on Oct 1 plus the maturity lag, so the first month-over-month comparison runs against a FULL prior month instead of a partial one.
- Monthly Config Watch #1: the same August window through the config search engine (w200), published as the fourth monthly report of this issue.
- Quarterly Stability Index: the next identical-prompt run lands around Oct 8 and is the first cross-quarter point on model stability.

Related research

Report	Direct PDF link
Monthly Calibration #1	https://marketmania.ai/research/reports/calibration-monthly-2026-08.pdf
Monthly Model Watch #1	https://marketmania.ai/research/reports/model-watch-monthly-2026-08.pdf
Consensus Watch #4	https://marketmania.ai/research/reports/consensus-watch-2026-08-24.pdf

The three monthly reports publish together as one issue each month; each links straight to the others' PDF and to the last weekly issue of the same series. Direct links are the posting rule from wave 2 on.

CITE THIS REPORT

```
@misc{mm_consensus_monthly_2026m08,
  title = {Monthly Consensus Watch #1: does agreeing with the crowd make an LLM's market call safer?},
  author = {{MarketMania Research}},
  year  = {2026}, month = {September}, day = {4},
  url   = {https://marketmania.ai/research/reports/consensus-watch-monthly-2026-08.pdf},
  note  = {Methodology v1.1; window Aug 1-31, 2026 UTC; source monthly_metrics_2026-08.json}
}
```