

Monthly Calibration #1

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

RESEARCH SNAPSHOT			
19,383 directional calls scored	7 models (lineage spliced: grok, qwen)	of 40,979 mature forecasts	window Aug 1-31, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 46.3% vs 62.6 stated	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)

Over the month the field's overconfidence gap read +16.3pp against 62.6 stated mean confidence, the weekly gaps ran +8.6pp to +20.4pp inside it, and the field slope across the four weeks is -1.21 pp/week (narrowing).

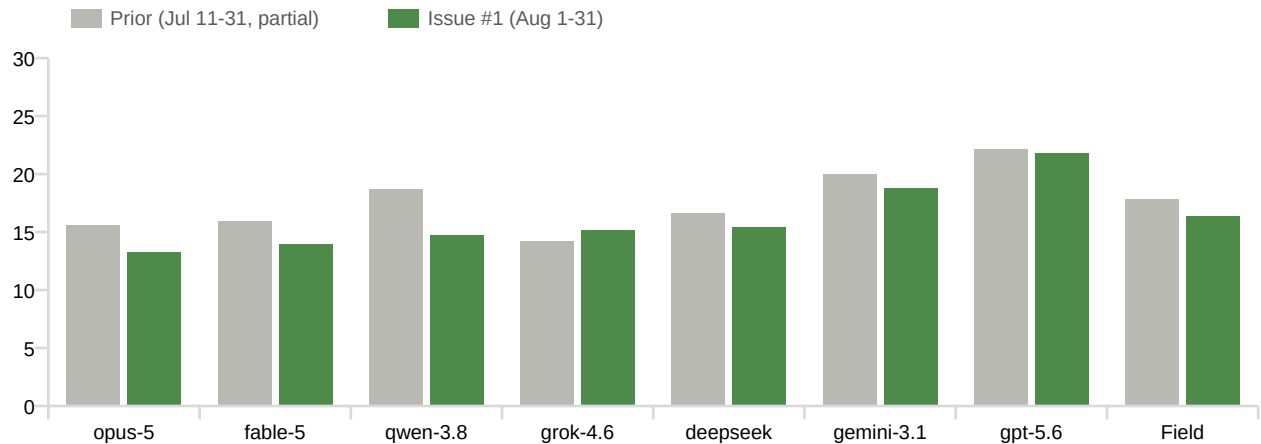
WEEKLY EVIDENCE -> MONTHLY VERDICT (rule below)

Metric	W1 Aug 3-9	W2 Aug 10-16	W3 Aug 17-23	W4 Aug 24-30	Month	Verdict
Field gap (conf - hit), pp	+20.4	+18.3	+8.6	+19.6	+16.3	Confirmed (4/4)
Lowest-Brier model	fable-5	qwen-3.8	opus-5	deepseek	opus-5	Rejected (1/4)
Widest-gap model	gpt-5.6	gpt-5.6	gpt-5.6	gpt-5.6	gpt-5.6	Confirmed (4/4)

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 -- The monthly gap read +16.3pp.** Directional calls hit 46.3% against 62.6 stated mean confidence over 19,383 calls. Field Brier is 0.2784 and 0 of 7 model lines scored below 0.25, the score of an uninformative always-50% predictor. One month, descriptive.
- **The field gap slope across the four weeks is -1.21 pp/week (narrowing).** Across those weeks the field gap ran +20.4pp -> +18.3pp -> +8.6pp -> +19.6pp and the per-model slopes split 5 narrowing / 0 widening / 2 flat under the pack's own rule ($|\text{slope}| < 0.5 \text{ pp/week} = \text{flat}$). claude-opus-5 is both the best-calibrated line (Brier 0.2677) and the hit-rate leader (47.8%).
- **The widest weekly swing belongs to claude-opus-5.** Its weekly gap moved over a 17.8 pp range inside one month, against a field range of 11.8 pp. Gap moves against the prior window ran from -4.0pp for qwen-3.8-max (spliced) to +0.9pp for grok-4.6 (spliced); 6 of 7 lines sit below their prior-window gap.
- **Trading, kept separate from every calibration table.** 0 of 7 lines finished net-positive over the month; net PnL ran -\$25.20 (claude-opus-5, best) to -\$243.84 (grok-4.6 (spliced), worst), field -\$1054.79.



Overconfidence gap in percentage points (mean stated confidence minus hit-rate), the prior window against the whole month, sorted by this month's gap. Short names: opus-5 = claude-opus-5, fable-5 = claude-fable-5, gemini-3.1 = gemini-3.1-pro, gpt-5.6 = gpt-5.6-sol, qwen-3.8 = qwen-3.8-max, deepseek = deepseek-v4-pro. Field gap +17.8pp -> +16.3pp; field mean stated confidence 62.8 -> 62.6. 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

Every MarketMania forecast carries a model-stated confidence from 0 to 100 alongside its direction call. This report checks whether that number tracks reality: on a well-calibrated forecaster, calls made at 70% confidence should hit their direction about 70% of the time. The weekly series carries one point per model per issue. A calendar month carries four weekly points plus a pooled monthly figure and a prior-window baseline, which is what the slope table and the week-by-week confidence buckets below are computed from. Descriptive for this window.

How to read this

Gap pp = mean stated confidence minus hit-rate, in percentage points; positive means overconfident. Brier is the mean squared error of the stated probability against the realised outcome, so lower is better and 0.25 is what an uninformative always-50% predictor scores. Coverage = calls scored divided by the mature calls available to that model. Confidence buckets are the models' own natural breakpoints, not an arbitrary binning; cells below N=10 are marked insufficient and never used to rank anything. Prediction and trading metrics live in separate sections and are never combined. 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)). The monthly rows are one pull at the monthly cutoff (Mon Sep 1, 2026, 16:00 UTC); the weekly rows are the published issues at their own frozen cutoffs, so the two do not have to agree to the last decimal.

Calibration by model (sorted by Brier, best first)

Model	N	Cov.	Hit rate	Mean conf	Gap pp	Brier	95% CI
claude-opus-5	2,281	38.9%	47.8%	61.0	+13.2	0.2677	45.7%-49.8%
claude-fable-5	2,812	48.0%	46.7%	60.6	+13.9	0.2689	44.9%-48.6%
grok-4.6 (spliced)	2,708	46.2%	45.3%	60.3	+15.1	0.2710	43.4%-47.1%
qwen-3.8-max (spliced)	3,001	51.7%	45.6%	60.3	+14.7	0.2730	43.9%-47.4%
deepseek-v4-pro	2,578	44.0%	45.5%	60.9	+15.4	0.2752	43.6%-47.5%
gemini-3.1-pro	3,033	51.8%	47.6%	66.4	+18.8	0.2881	45.8%-49.4%
gpt-5.6-sol	2,970	50.6%	45.9%	67.7	+21.8	0.3005	44.1%-47.6%
Field (all models)	19,383	-	46.3%	62.6	+16.3	0.2784	-

N = directional calls scored this window. Cov. = coverage, scored / mature calls available to that model. Gap pp = mean stated confidence minus hit-rate (positive = overconfident). 95% CI is the descriptive Wilson interval on hit-rate. Prior gaps per model are

the gray bars on page 1: prior window Jul 11-31 (partial), field gap +17.8pp against +16.3pp here, on n=19,541 directional calls there. 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).

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
claude-fable-5	46.1% (900)	47.0% (1,912)	n/a (0)	n/a (0)
claude-opus-5	49.4% (528)	47.3% (1,753)	n/a (0)	n/a (0)
deepseek-v4-pro	44.7% (853)	46.4% (1,566)	40.0% (140)	25.0% (4) *
gemini-3.1-pro	48.0% (25)	48.0% (1,953)	46.5% (1,028)	59.3% (27)
gpt-5.6-sol	37.5% (16)	48.1% (2,078)	40.7% (870)	50.0% (6) *
grok-4.6 (spliced)	44.3% (1,324)	46.1% (1,370)	50.0% (14)	n/a (0)
qwen-3.8-max (spliced)	47.6% (1,306)	44.4% (1,488)	44.1% (161)	20.0% (5) *

n in parentheses. * = N<10 (insufficient) -- no conclusions drawn from these cells. "n/a (0)" = no calls landed in that bucket this month. Sub-50 bucket (n greater than 0 only): qwen-3.8-max (spliced) 36.6% (n=41); deepseek-v4-pro 66.7% (n=15). Mid-scale ranking: 5 of the 7 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this month.

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	46.5% (1,776)	46.2% (861)	51.4% (175)
claude-opus-5	47.2% (1,357)	47.7% (761)	52.8% (163)
deepseek-v4-pro	45.1% (1,576)	46.3% (851)	46.4% (151)
gemini-3.1-pro	47.4% (1,951)	46.6% (912)	55.3% (170)
gpt-5.6-sol	46.2% (1,850)	45.1% (956)	46.3% (164)
grok-4.6 (spliced)	45.1% (1,713)	45.1% (851)	47.9% (144)
qwen-3.8-max (spliced)	45.4% (1,887)	46.0% (933)	46.4% (181)

n in parentheses; * = N<10 (insufficient). 7 of 7 model lines hit better on 1d than on 1h this month.

Trading result (a different question)

A different question: is any of this profitable to trade. "Accurate" and "profitable" are not the same thing, and per methodology this section is never mixed into the prediction/calibration metrics above.

Model	Trades	WR	Net PnL	Gross PnL	Max DD
claude-opus-5	2,281	37.9%	-\$25.20	+\$202.90	-\$160.92
claude-fable-5	2,812	36.2%	-\$96.55	+\$184.65	-\$210.11
gemini-3.1-pro	3,033	36.3%	-\$98.64	+\$204.66	-\$197.07
deepseek-v4-pro	2,578	36.6%	-\$176.06	+\$81.74	-\$182.97
qwen-3.8-max (spliced)	3,001	35.0%	-\$198.93	+\$101.17	-\$231.65
gpt-5.6-sol	2,970	35.2%	-\$215.58	+\$81.42	-\$235.74
grok-4.6 (spliced)	2,708	33.5%	-\$243.84	+\$26.96	-\$247.38

0 of 7 model lines finished net-positive and 7 gross-positive over the month, against 0 of 7 net-positive in the prior window Jul 11-31 (partial) on a summed net PnL of -\$3095.02. Ranked best-to-worst net PnL. Per methodology this table is never combined with the calibration tables above. The pack labels two of the four weeks calm and two volatile.

Platform note: TP/SL calibration

Context, not a finding. Since Aug 13 the public sandbox can trade with model-specific TP/SL multipliers learned from this same history (v1); since Aug 19, v2 adds per-ticker and confidence-bucket (50-70 / 70-100) resolution. Both switches landed INSIDE this window, so the month spans three feature states. That feature consumes calibration history; it does not feed back into any table in this report, which measures stated confidence vs. direction-hit only. No effect size is claimed for v1 or v2. The toggle lives at marketmania.ai/indices.

Overconfidence drift, week by week

The four Mon-Sun weeks of August are the same four published weekly issues, so the monthly gap splits into four weekly gaps of +20.4pp -> +18.3pp -> +8.6pp -> +19.6pp. Slope is an ordinary least-squares fit over weeks 1..4; the pack calls a line "flat" when |slope| is under 0.5 pp/week. 5 of 7 lines read narrowing, 0 widening and 2 flat, and the field line reads narrowing at -1.21 pp/week.

Model	W1 gap	W2 gap	W3 gap	W4 gap	Month gap	Slope pp/wk	Trend
claude-opus-5	+20.0	+19.2	+2.2	+18.8	+13.2	-2.06	narrowing
claude-fable-5	+19.2	+18.1	+4.7	+17.9	+13.9	-1.73	narrowing
qwen-3.8-max (spliced)	+18.3	+14.1	+7.4	+16.9	+14.7	-1.09	narrowing
grok-4.6 (spliced)	+18.3	+15.8	+9.4	+20.3	+15.1	-0.04	flat
deepseek-v4-pro	+20.3	+18.2	+9.4	+15.9	+15.4	-2.20	narrowing
gemini-3.1-pro	+22.2	+18.4	+13.1	+22.5	+18.8	-0.44	flat
gpt-5.6-sol	+24.4	+24.0	+13.9	+24.8	+21.8	-0.89	narrowing
Field (all models)	+20.4	+18.3	+8.6	+19.6	+16.3	-1.21	narrowing

Weeks are W1 Aug 3-9, W2 Aug 10-16, W3 Aug 17-23, W4 Aug 24-30. Slope = OLS over weeks 1..4, in pp per week; "flat" means |slope| < 0.5 pp/week. Source: the frozen weekly packs, with the 7-line roster and the lineage splices applied, so for weeks 1-3 the published PDFs carried the then-current roster and small differences to those PDFs are expected. The Month gap is the pooled month, not the mean of the four weeks; the prior window Jul 11-31 (partial) read a field gap of +17.8pp. 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).

Bucket	W1	W2	W3	W4
50-60	43.8% vs 57.0 (985)	44.3% vs 56.9 (1,222)	50.2% vs 57.2 (1,182)	45.5% vs 57.0 (1,079)
60-70	42.9% vs 63.2 (2,528)	43.7% vs 63.1 (2,021)	55.5% vs 63.0 (3,411)	42.6% vs 62.9 (2,994)
70-80	38.0% vs 72.8 (498)	42.1% vs 73.0 (406)	53.1% vs 72.8 (518)	39.2% vs 73.0 (485)
80-100	33.3% vs 81.2 (9) *	20.0% vs 80.4 (5) *	84.6% vs 81.2 (13)	41.7% vs 80.6 (12)

Field level: every model's calls pooled into the stated-confidence bucket. Cells are hit rate vs mean stated confidence, with n in parentheses; * = N<10 (insufficient), "n/a (0)" = no call landed in that bucket that week. Source: the month pull split by slot week at the monthly cutoff, so the four columns are one consistent pool.

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

- Pooled over a whole month the field's gap is +16.3pp on 19,383 scored calls: stated confidence averaged 62.6 while the direction hit 46.3%.
- The field gap spans 11.8 pp across the four weeks inside the month, against a pooled monthly gap of +16.3pp and a prior-window gap of +17.8pp; field mean stated confidence moved 62.8 -> 62.6 over the same pair of windows.
- Trading this month finished 0 of 7 lines net-positive against 0 of 7 in the prior window; the pack labels two of the four weeks calm and two volatile.

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.
- Prediction metrics (this report) and trading metrics are kept in separate sections per methodology -- they are never combined into a single score. 95% Wilson CIs shown are descriptive, not inferential: observations inside one window are dependent, so read them as a range, not a formal coverage guarantee.
- All 19,383 scored calls sit inside one calendar month, and four weekly points are four points: an OLS slope over them is a description of this month, not evidence of drift. 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. Models report confidence at discrete levels, not a continuous scale; the buckets reflect those natural breakpoints and cells below N=10 are marked insufficient.
- Underlying price series are reconstructed from trade entry prices (median per symbol-slot), not an independent tick feed. 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: 19,383 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 · 20 published reports.

Issue #1. Monthly Calibration is a living monthly comparison; each issue appends one more calendar month of confidence-vs-hit data. This first issue sets the baseline: field gap +16.3pp on 19,383 scored calls, field Brier 0.2784, weekly gaps +20.4pp -> +18.3pp -> +8.6pp -> +19.6pp, field slope -1.21 pp/week (narrowing). 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 Consensus Watch #1	https://marketmania.ai/research/reports/consensus-watch-monthly-2026-08.pdf
Monthly Model Watch #1	https://marketmania.ai/research/reports/model-watch-monthly-2026-08.pdf
Weekly Calibration #4	https://marketmania.ai/research/reports/weekly-calibration-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_calibration_monthly_2026m08,
  title = {Monthly Calibration #1: confidence vs. direction-hit, Aug 1-31 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {4},
  url = {https://marketmania.ai/research/reports/calibration-monthly-2026-08.pdf},
  note = {Methodology v1.1, hash e66c7e8c864a2233; source monthly_metrics_2026-08.json}
}
```