

Monthly Calibration #2

October 2, 2026 · MarketMania Research · Monthly series · window Sep 1-30, 2026 UTC

RESEARCH SNAPSHOT

20,682 directional calls scored	7 models (stable lineup)	of 39,806 mature forecasts	window Sep 1-30, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 45.5% vs 62.6 stated	cutoff Thu Oct 1, 2026, 16:00 UTC
market: BTC net +6.42% (prior +24.95%)	ann vol 41.2% (was 41.5%)	TOP5 volume \$80.01B, +21% m/m	pairwise corr 0.78 (was 0.79)

KEY FINDING · OBSERVATION (one monthly window)

Over the month the field's overconfidence gap read +17.0pp against 62.6 stated mean confidence, the weekly gaps ran +12.1pp to +22.5pp inside it, and the field slope across the four weeks is -2.09 pp/week (narrowing).

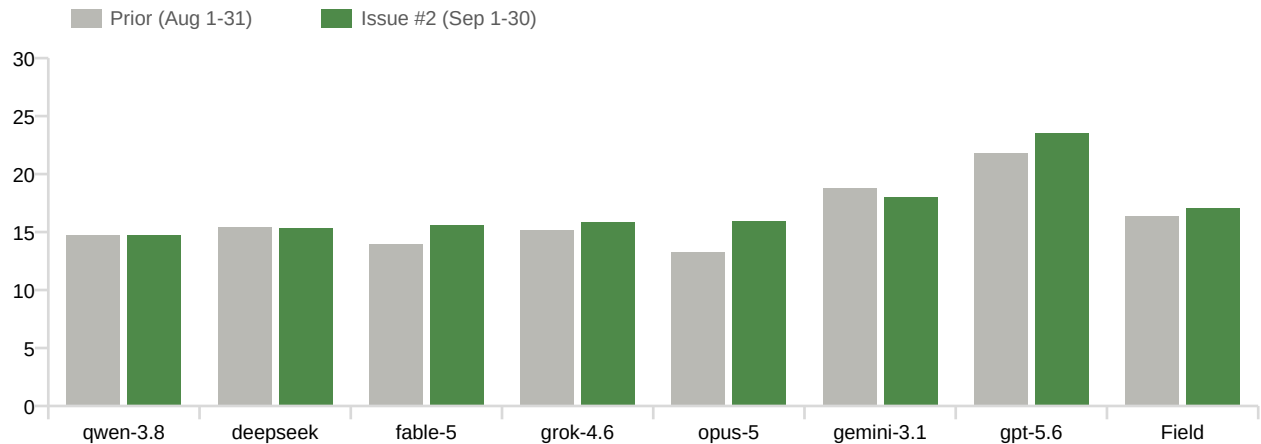
WEEKLY EVIDENCE -> MONTHLY VERDICT (rule below)

Metric	W1 Aug 31-Sep 6	W2 Sep 7-13	W3 Sep 14-20	W4 Sep 21-27	Month	Verdict
Field gap (conf - hit), pp	+19.6	+22.5	+12.1	+16.1	+17.0	Confirmed (4/4)
Lowest-Brier model	qwen-3.8	qwen-3.8	fable-5	qwen-3.8	qwen-3.8	Confirmed (3/4)
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); the month is the pooled Sep 1-30 pull. W1 straddles the month edge (Aug 31).

TL;DR

- **OBSERVATION -- The monthly gap read +17.0pp.** Directional calls hit 45.5% against 62.6 stated mean confidence over 20,682 calls. Field Brier is 0.2796 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 -2.09 pp/week (narrowing).** Across those weeks the field gap ran +19.6pp -> +22.5pp -> +12.1pp -> +16.1pp and the per-model slopes split 6 narrowing / 0 widening / 1 flat under the pack's own rule ($|\text{slope}| < 0.5 \text{ pp/week} = \text{flat}$). qwen-3.8-max is the lowest-Brier line (Brier 0.2701); gemini-3.1-pro leads on hit-rate (48.5%).
- **The widest weekly swing belongs to claude-opus-5.** Its weekly gap moved over a 14.6 pp range inside one month, against a field range of 10.4 pp. Gap moves against the prior window ran from -0.8pp for gemini-3.1-pro to +2.7pp for claude-opus-5; 2 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 -\$305.20 (gemini-3.1-pro, best) to -\$440.88 (deepseek-v4-pro, worst), field -\$2554.15.



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 +16.3pp -> +17.0pp; field mean stated confidence 62.6 -> 62.6. The prior window is the full August: window A of the same pull, slots 2026-08-01 to 2026-09-01 (first slot actually seen 2026-08-01 00:01 UTC), matured at the prior cutoff Tue Sep 1, 2026, 16:00 UTC under the same engine, the same FH gate and the same hit rule as September -- 19,383 directional calls against 20,682 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. No model line is lineage-spliced this month (the prior window carries the grok and qwen splices). The monthly rows are one pull at the monthly cutoff (Thu Oct 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
qwen-3.8-max	3,165	56.0%	44.7%	59.5	+14.7	0.2701	43.0%-46.5%
deepseek-v4-pro	3,352	59.1%	45.3%	60.5	+15.3	0.2729	43.6%-46.9%
claude-fable-5	2,945	51.7%	45.4%	61.0	+15.6	0.2733	43.6%-47.2%
claude-opus-5	2,532	44.5%	45.4%	61.3	+15.9	0.2740	43.5%-47.3%
grok-4.6	2,271	40.0%	44.4%	60.1	+15.8	0.2740	42.4%-46.4%
gemini-3.1-pro	3,284	57.6%	48.5%	66.5	+18.0	0.2848	46.8%-50.2%
gpt-5.6-sol	3,133	55.0%	44.5%	68.0	+23.5	0.3054	42.7%-46.2%
Field (all models)	20,682	-	45.5%	62.6	+17.0	0.2796	-

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 Aug 1-31, field gap +16.3pp against +17.0pp here, on n=19,383 directional calls there. In the prior window the lineage splice folds 4,520 grok-4.5 rows into grok-4.6 and 870 qwen-3.7-max rows into qwen-3.8-max (5,390 rows in all).

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
claude-fable-5	43.6% (866)	46.1% (2,078)	0.0% (1) *	n/a (0)
claude-opus-5	46.8% (462)	45.1% (2,070)	n/a (0)	n/a (0)
deepseek-v4-pro	47.0% (1,193)	44.2% (2,075)	46.3% (82)	n/a (0)
gemini-3.1-pro	27.8% (18)	48.9% (2,097)	48.0% (1,140)	50.0% (28)
gpt-5.6-sol	55.2% (29)	45.1% (2,079)	43.1% (1,019)	16.7% (6) *
grok-4.6	47.8% (837)	42.4% (1,431)	0.0% (2) *	n/a (0)
qwen-3.8-max	44.3% (1,511)	45.4% (1,602)	31.2% (16)	n/a (0)

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 38.9% (n=36); deepseek-v4-pro 50.0% (n=2) *; grok-4.6 100.0% (n=1) *; gemini-3.1-pro 100.0% (n=1) *. Mid-scale ranking: 3 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	47.5% (1,870)	42.6% (903)	36.0% (172)
claude-opus-5	47.7% (1,551)	43.2% (833)	33.1% (148)
deepseek-v4-pro	47.3% (2,095)	43.2% (1,042)	35.4% (215)
gemini-3.1-pro	50.1% (2,132)	47.1% (965)	37.4% (187)
gpt-5.6-sol	45.8% (1,940)	43.4% (1,009)	36.4% (184)
grok-4.6	46.4% (1,436)	42.2% (711)	33.9% (124)
qwen-3.8-max	46.3% (1,984)	43.0% (998)	36.6% (183)

n in parentheses; * = N<10 (insufficient). 0 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
gemini-3.1-pro	3,284	40.1%	-\$305.20	+\$23.20	-\$318.13
grok-4.6	2,271	37.5%	-\$308.62	-\$81.52	-\$313.05
claude-opus-5	2,532	37.9%	-\$331.35	-\$78.15	-\$337.15
claude-fable-5	2,945	38.4%	-\$348.13	-\$53.63	-\$353.47
qwen-3.8-max	3,165	38.0%	-\$397.69	-\$81.19	-\$401.28
gpt-5.6-sol	3,133	37.6%	-\$422.29	-\$108.99	-\$428.47
deepseek-v4-pro	3,352	38.8%	-\$440.88	-\$105.68	-\$443.74

0 of 7 model lines finished net-positive and 1 gross-positive over the month, against 0 of 7 net-positive in the prior window Aug 1-31 on a summed net PnL of -\$1054.80. 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 before this window, inside the prior window (Aug 13 and Aug 19). 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 from Aug 31 to Sep 27 are the published weekly issues #5, #6, #7 and #8, and their weekly gaps read +19.6pp -> +22.5pp -> +12.1pp -> +16.1pp. 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. 6 of 7 lines read narrowing, 0 widening and 1 flat, and the field line reads narrowing at -2.09 pp/week.

Model	W1 gap	W2 gap	W3 gap	W4 gap	Month gap	Slope pp/wk	Trend
qwen-3.8-max	+16.6	+18.8	+9.2	+15.6	+14.7	-1.26	narrowing
deepseek-v4-pro	+16.5	+18.4	+12.4	+17.0	+15.3	-0.45	flat
claude-fable-5	+17.9	+23.6	+9.1	+14.7	+15.6	-2.41	narrowing
grok-4.6	+22.3	+20.5	+10.7	+13.5	+15.8	-3.62	narrowing
claude-opus-5	+18.8	+24.5	+9.9	+14.0	+15.9	-2.90	narrowing
gemini-3.1-pro	+19.7	+23.6	+14.0	+16.2	+18.0	-2.01	narrowing
gpt-5.6-sol	+26.2	+28.4	+18.4	+21.0	+23.5	-2.56	narrowing
Field (all models)	+19.6	+22.5	+12.1	+16.1	+17.0	-2.09	narrowing

Weeks are W1 Aug 31-Sep 6, W2 Sep 7-13, W3 Sep 14-20, W4 Sep 21-27. Slope = OLS over weeks 1..4, in pp per week; "flat" means |slope| < 0.5 pp/week. Source: the frozen weekly packs (7-line roster), i.e. the windows of the published weekly issues #5, #6, #7 and #8 at their own cutoffs. The Month gap is the pooled month, not the mean of the four weeks; the prior window Aug 1-31 read a field gap of +16.3pp. In the prior window the lineage splice folds 4,520 grok-4.5 rows into grok-4.6 and 870 qwen-3.7-max rows into qwen-3.8-max (5,390 rows in all).

Bucket	W1	W2	W3	W4
50-60	47.6% vs 56.9 (900)	43.2% vs 56.9 (1,261)	50.6% vs 57.0 (1,166)	42.8% vs 57.0 (1,037)
60-70	43.6% vs 62.9 (2,500)	37.8% vs 62.7 (2,719)	51.1% vs 62.9 (3,779)	47.4% vs 63.1 (3,319)
70-80	42.4% vs 72.7 (417)	39.8% vs 73.1 (472)	48.9% vs 72.7 (644)	51.3% vs 73.0 (575)
80-100	57.1% vs 81.4 (7) *	0.0% vs 82.0 (1) *	37.5% vs 80.8 (8) *	47.1% vs 80.0 (17)

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. W1 holds Sep 1-6 here: the month pull starts on Sep 1.

Market check: September ran calmer on vol, wider on price

OBSERVATION — mean |1d move| 1.98% -> 1.91% (-4% rel), BTC realized vol 36.9% -> 34.9% (ann., hourly); field directional accuracy 46.3% -> 45.5% (-0.8pp), 1 of 7 models improved, sim win-rate up for 6 of 7, field sim PnL -\$1054.79 -> -\$2554.15 (Aug 1-31 vs Sep 1-30; descriptive, one pair of windows, not a claim).

Robustness: raw price-sign accuracy 46.0% -> 48.4% — the opposite direction to the trade-based hit rule.

Measure	Aug 1-31	Sep 1-30	Change
Mean 1d move (5 assets)	1.98%	1.91%	-4% rel
BTC realized vol (ann., hourly)	36.9%	34.9%	-2.0pp
Field directional accuracy	46.3%	45.5%	-0.8pp
Models improving hit-rate	--	1 of 7	--
Raw price-sign accuracy	46.0%	48.4%	+2.4pp
Field sim win-rate	35.8%	38.4%	+2.6pp
Field sim net PnL	-\$1054.79	-\$2554.15	--

Window A is the pack's prior window -- the full August -- so this slice and every month-over-month column in this report compare September against the full August. 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: none. 0 rows were remapped inside the month against 5,390 in the prior window.

Practical implications

- Pooled over a whole month the field's gap is +17.0pp on 20,682 scored calls: stated confidence averaged 62.6 while the direction hit 45.5%.
- The field gap spans 10.4 pp across the four weeks set against the month, against a pooled monthly gap of +17.0pp and a prior-window gap of +16.3pp; field mean stated confidence moved 62.6 → 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 the full August: every "prior" column, the gray series of the page-1 chart and the market-check A column describe Aug 1-31. It is window A of the same pull, slots 2026-08-01 to 2026-09-01 (first slot actually seen 2026-08-01 00:01 UTC), matured at the prior cutoff Tue Sep 1, 2026, 16:00 UTC under the same engine, the same FH gate and the same hit rule as September -- 19,383 directional calls against 20,682 in the month. Month-over-month deltas are a step between two windows of different length (31 and 30 days).
- Two cutoffs live in this issue. The month is frozen at Thu Oct 1, 2026, 16:00 UTC; the four weekly blocks are frozen at their own published cutoffs (W1 Aug 31-Sep 6, W2 Sep 7-13, W3 Sep 14-20, W4 Sep 21-27). 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 (20,682 vs 19,544 directional, plus 1,826 on the 3 remainder days); W1 also opens on Aug 31, a day the month pull does not hold, so its pack counts 4,520 directional calls against 3,832 for Sep 1-6 inside the month pull -- 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 20,682 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 10 trend days of 30, and the pack's own volatility rule splits the four Mon-Sun weeks from Aug 31 to Sep 27 as W1 volatile, W2 calm, W3 volatile, W4 calm; 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 #2): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so across this 30-day window daily coverage is FULL as measured -- the 1w series has slots on 30 of 30 days and the 1M series on 30 of 30 days; both sit outside this report's FH gate (1h/4h/1d) and appear only in the exclusions counter (1w 2,096, 1M 2,089). (2) No model line is lineage-spliced inside this window. (3) Every per-model aggregate in this issue sees the merged head id; the raw id is kept in the pack, and 0 rows were remapped inside the month against 5,390 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 the full August: it covers 31 days (Aug 1-31), and the TOP5 volume figure compares a 30-day sum with a 31-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 (Thu Oct 1, 2026, 16:00 UTC) -- the D2 definition pinned in the weekly series; this issue prints 59,755. 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 (Sep 21-27) reproduces the published weekly #8 on 7 of 7 checks, and the prior window of this pack (Aug 1-31, prior cutoff Tue Sep 1, 2026, 16:00 UTC) reproduces the published Monthly #1 on 7 of 7 checks. Counter deltas against pre-wave-3 issues remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	44,100	Mature (scored pool)	39,806
OK in gate	39,806	-- of them directional	20,682
Out of gate (1w / 1M)	2,096 / 2,089	-- of them sideways	19,124
Invalid	109	Pending (next issue)	0
Late closes	0	Uptime, all grids	1,140 / 1,140 slots
Source file	monthly_metrics_2026-09.json	Report cutoff	Thu Oct 1, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-02 08:05 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 20,682 scored observations. MARKETMANIA RESEARCH TO DATE (as of cutoff Oct 1, 16:00 UTC): 59,755 resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 40 published reports.

Issue #2. Monthly Calibration is a living monthly comparison; each issue appends one more calendar month of confidence-vs-hit data. This second issue reads September against the full August: field gap +17.0pp on 20,682 scored calls, field Brier 0.2796, weekly gaps +19.6pp -> +22.5pp -> +12.1pp -> +16.1pp, field slope -2.09 pp/week (narrowing), against a prior-window field gap of +16.3pp. No model line is lineage-spliced this month (the prior window carries the grok and qwen splices). Engine 1.1 has powered the sandbox since Aug 18, i.e. before this window and inside the prior window; no cross-engine PnL comparisons are claimed.

What we're testing next

- Monthly #3 (October): the same three series over Oct 1-31, cut on Nov 1 plus the maturity lag, so its month-over-month comparison runs against this September.
- Monthly Config Watch #2: the same September window through the config search engine (w298), published as the fourth monthly report of this issue.
- Quarterly Stability Index: Run 3, the next repeat-prompt stability run, is planned for the first half of October and is the first cross-quarter point on model stability.

Related research

Report	Direct PDF link
Monthly Consensus Watch #2	https://marketmania.ai/research/reports/consensus-watch-monthly-2026-09.pdf
Monthly Model Watch #2	https://marketmania.ai/research/reports/model-watch-monthly-2026-09.pdf
Weekly Calibration #8	https://marketmania.ai/research/reports/weekly-calibration-2026-09-21.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_2026m09,
  title = {Monthly Calibration #2: confidence vs. direction-hit, Sep 1-30 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {October}, day = {2},
  url = {https://marketmania.ai/research/reports/calibration-monthly-2026-09.pdf},
  note = {Methodology v1.1, hash e66c7e8c864a2233; source monthly_metrics_2026-09.json}
}
```